Данные как продукт:
зачем покрывать DWH метриками и что можно из этого получить

Ермаков Евгений, Яндекс Go



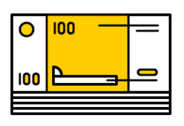


Что такое Яндекс Go?



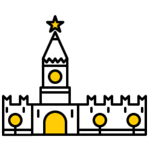
700 тыс.

активных водителей, которые сделали более одного заказа в месяц



18

стран присутствия, в том числе Гана и Кот-д'Ивуар



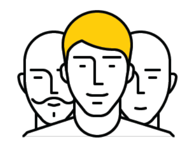
1000

городов, из них 300 крупных

За 2019



Хранилище данных Яндекс Go



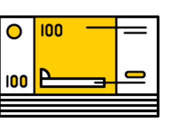
>500

уникальных пользователей данных в месяц



>900

отчетов по различным тематикам



4

крупных бизнес-юнита: Такси, Еда, Лавка, Драйв



3 Пб

накопленных данных по четырем бизнес-юнитам



Метрики хранилища

DMP - Usage KPI - All teams

Datasource updated at 09.05.2021 8:57:02

Report period	Report date	DAU/CNT	System	Robots	DMP activity	Team
Last 13 months	18.04.2021	CNT	(All)	False	False	All teams

Usage by layer - All teams

Select Layer to drill down

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

	Production	Not Production
cdm	19,38%	0,37%
dds	0,98%	0,27%
ods	21,86%	0,24%
raw	8,61%	0,40%
rep	0,02%	0,00%
snb	0,00%	44,25%
tech	3,44%	0,17%
udm	0,02%	0,00%

Production Kpi

Select KPI to filter

25,6%

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

54,31%
▲4,4% MoM
GP 15,75%
YT 38,55%

Team Kpi - All teams

Select KPI to filter

CDM / REP or Summary

7,3%

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

19,77%
▲0,1% MoM

Not Deprecated

76,7%

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

79,34%
▲0,7% MoM

Not Tech (RAW / STG)

90,0%

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

87,38%
▼1,6% MoM

Production by team

Select Team to filter

All teams
Atlas
Demand
Marketplace
MassService
Supply Eda
Supply Taxi
Undefined

Production - Object

Select Object to drill down

Object / Dep / User: Object
DAU/CNT at least: 200

YT	dm_order cdm [dep]	758 697
YT	quality_control_log raw	206 596
YT	driver_session_reduced dds [dep]	187 349
YT	dim_driver dds [dep]	146 763
YT	pin_stats raw	142 937
YT	dm_order cdm	136 843
YT	user_phone ods	112 053
YT	order_proc raw	106 061

Not Production - Object

YT	disinfected_cars_table_ snb	120 584
GP	voytekh_manager_v2 snb [dep]	115 624
GP	voytekh_pd_user snb [dep]	102 287
YT	accra_branding_subs snb	100 995
YT	helsinki_week snb	100 847
YT	support_tickets cdm [dep]	96 051



О ЧЕМ ДОКЛАД?

I. ЗАЧЕМ DWH МЕТРИКИ?
II. КАК РЕАЛИЗОВАЛИ?
III. ЧТО ПОЛУЧИЛИ?
IV. СТОИЛО ЛИ ТОГО?



I. ЗАЧЕМ DWH МЕТРИКИ?

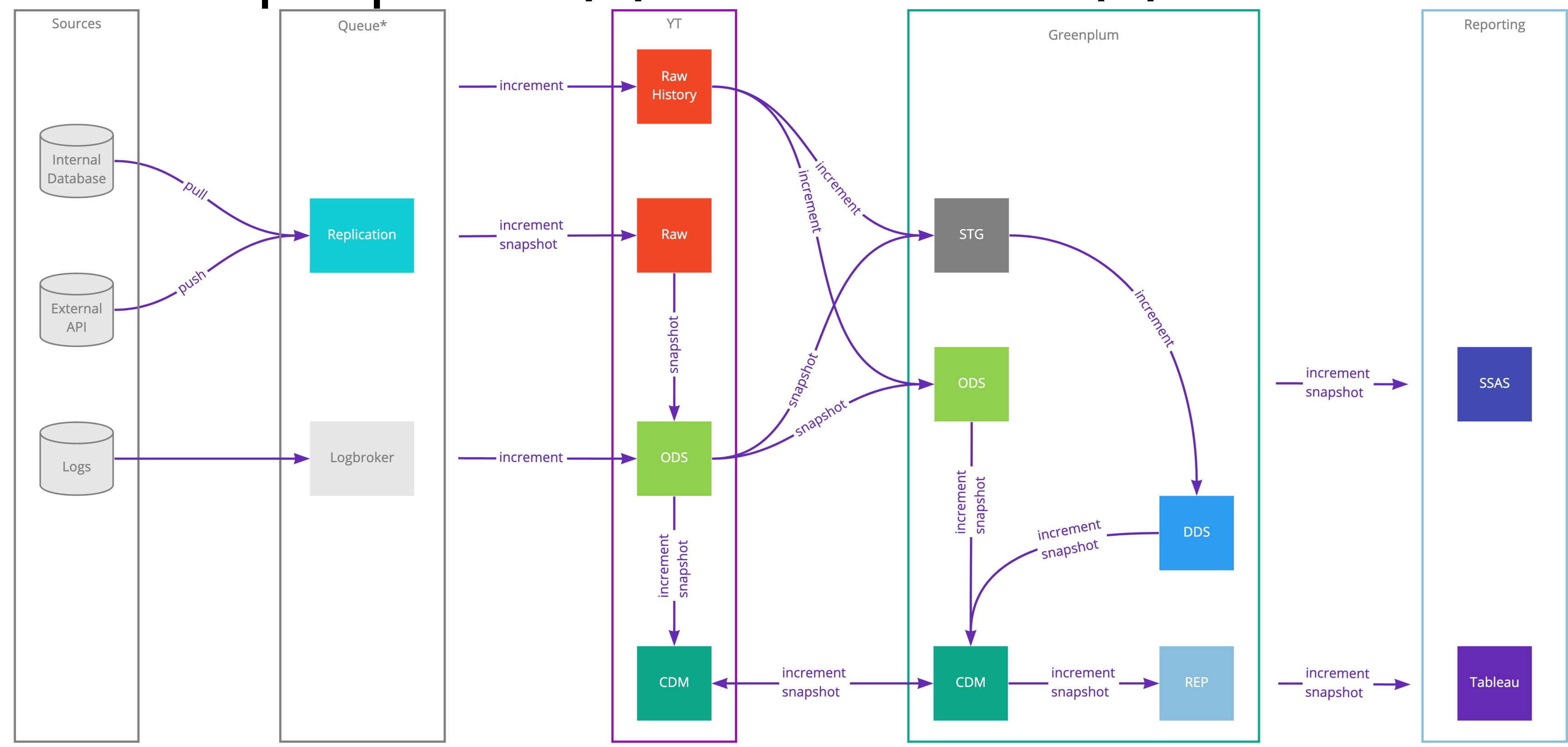
II. КАК РЕАЛИЗОВАЛИ?

III. ЧТО ПОЛУЧИЛИ?

IV. СТОИЛО ЛИ ТОГО?



Платформа данных Яндекс GO



 Владимир Верстов, https://smartdataconf.ru/2020/spb/talks/3fnm49atnbkyknqemzbtmc/

I. Зачем DWH метрики?

ПОЧЕМУ ТАК СЛОЖНО?

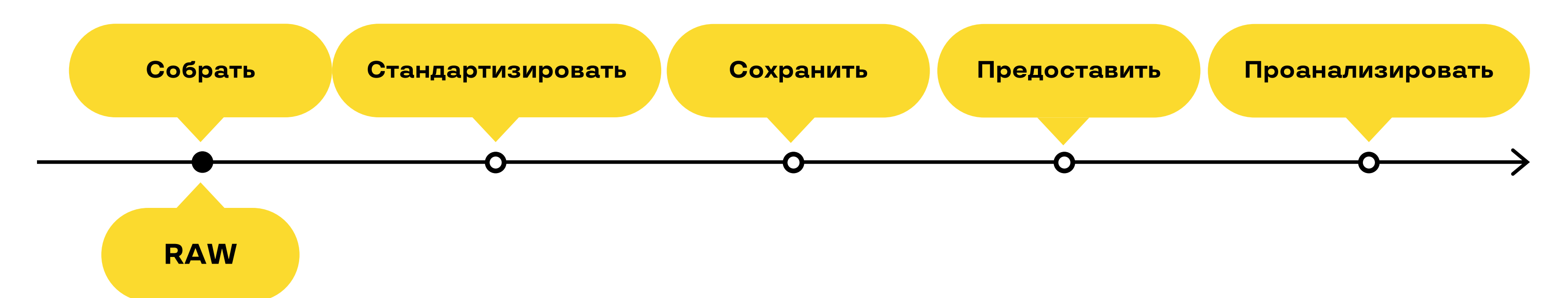


Архитектура слоев данных





Архитектура слоев данных



Цель

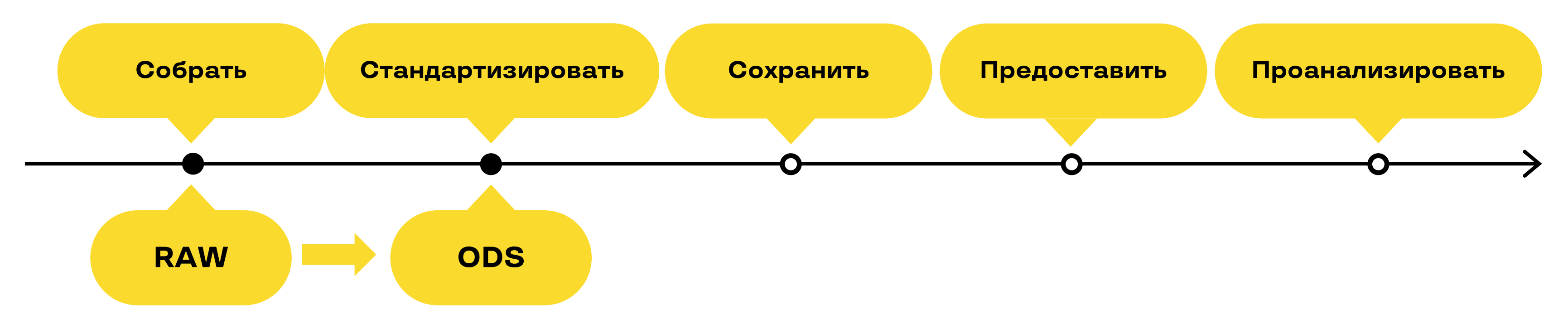
- Захватить сигналы источника

Задачи

- Собрать данные с источника as-is
- Преобразовать их в объекты с понятным описанием и методом доступа



Архитектура слоев данных



Цель

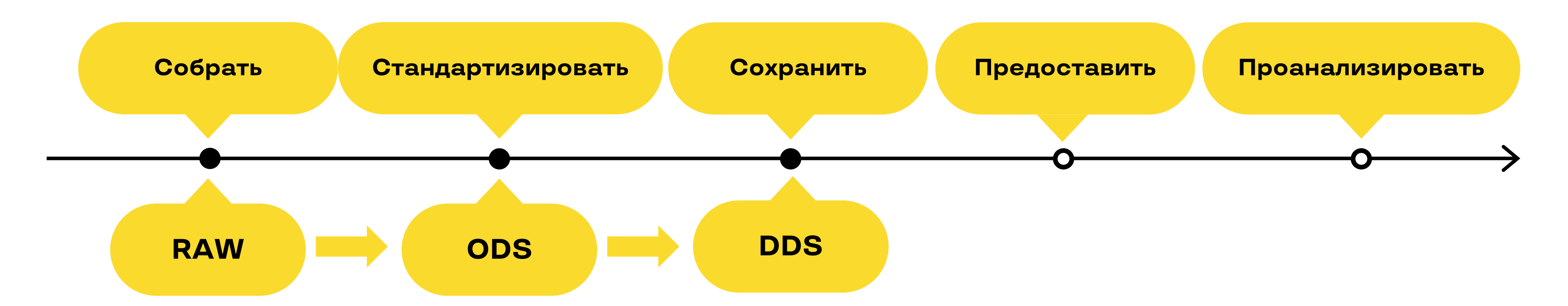
- Хранить операционные данные источника

Задачи

- Сформировать набор сущностей источника
- Разложить данные по сущностям
- Предоставить стандартный интерфейс доступа к данным



Архитектура слоев данных



Цель

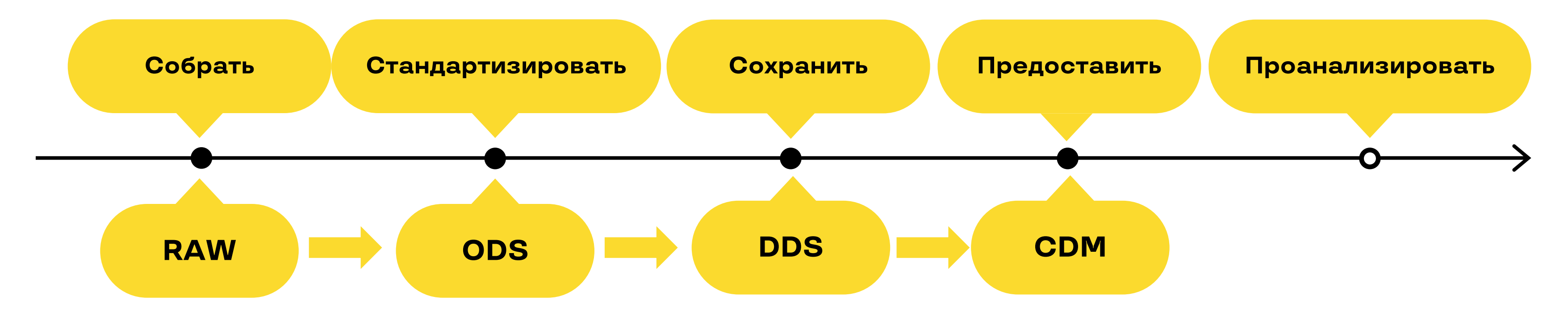
- Накапливать данные о сущностях доменной модели

Задачи

- Хранить детальную историю изменений
- Консолидировать данные между источниками



Архитектура слоев данных



Цель

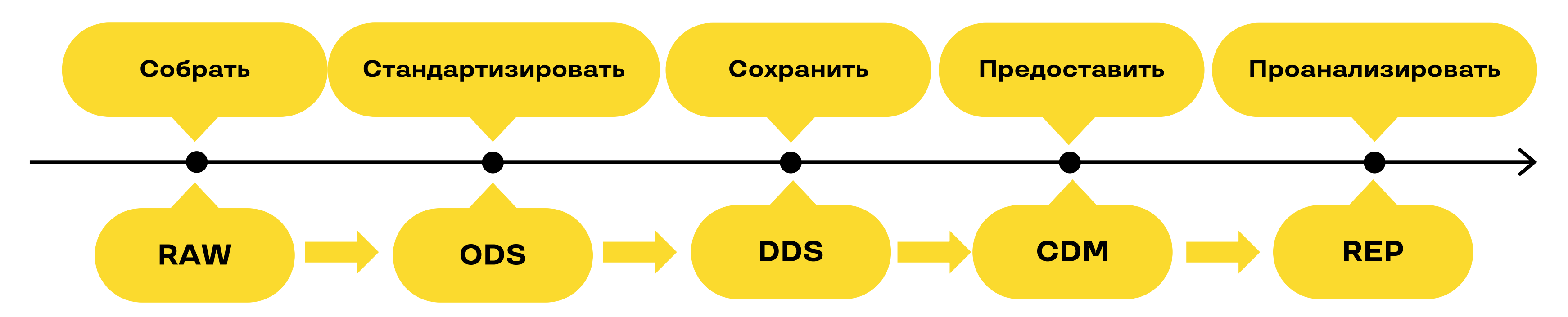
- Предоставлять витрины данных для анализа

Задачи

- Формировать данные в контексте бизнес-потребностей
- Оптимизировать доступ на чтение



Архитектура слоев данных



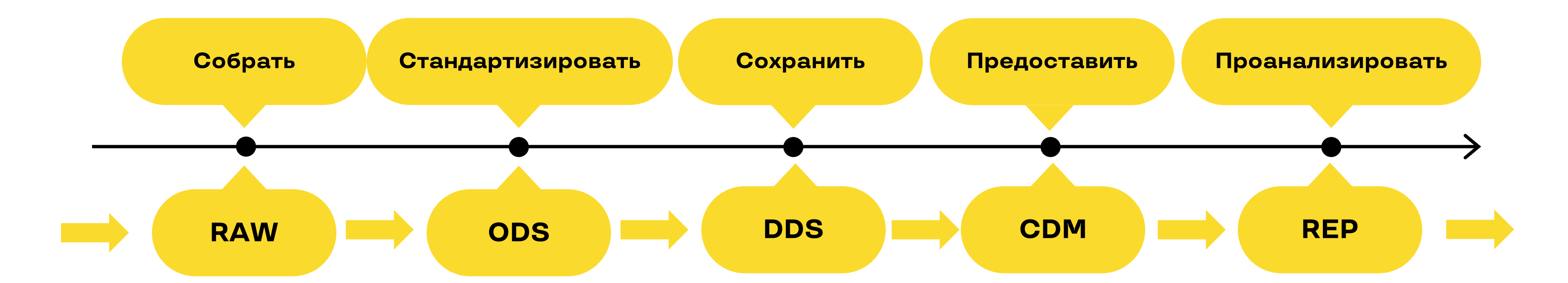
Цель

- Хранить отчетные срезы

Задачи

- Формировать данные в контексте бизнес-потребностей
- Готовить агрегированные отчеты

Архитектура слоев данных





I. Зачем DWH метрики?

КАК СЛОИ СВЯЗАНЫ С СИСТЕМАМИ?



Архитектура слоев данных

Собрать — RAW

Стандартизировать — ODS

Сохранить — DDS

Предоставить — CDM

Проанализировать — REP

Source Queue

- Забирает инкременты и снепшоты с источников различных типов
- Преобразовывает данные в устойчивый к изменениям формат



Архитектура слоев данных

Собрать → RAW
Стандартизировать → ODS
Сохранить → DDS
Предоставить → CDM
Проанализировать → REP

Source Queue

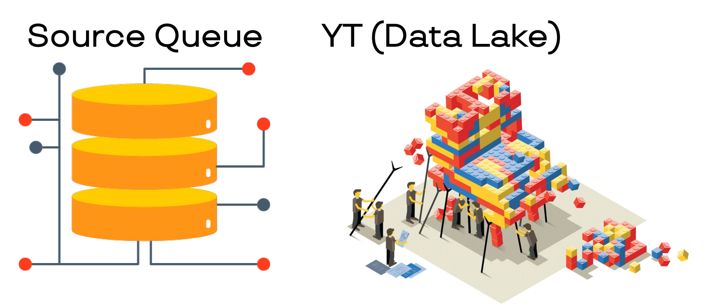
YT (Data Lake)

- Полуструктурированные данные
- Каркас MapReduce
- Аналоги экосистемы hadoop



Архитектура слоев данных

Собрать → RAW
Стандартизировать → ODS
Сохранить → DDS
Предоставить → CDM
Проанализировать → REP



Greenplum (Data warehouse)

- Различные ad-hoc-запросы
- Большое количество join
- Малое время отклика

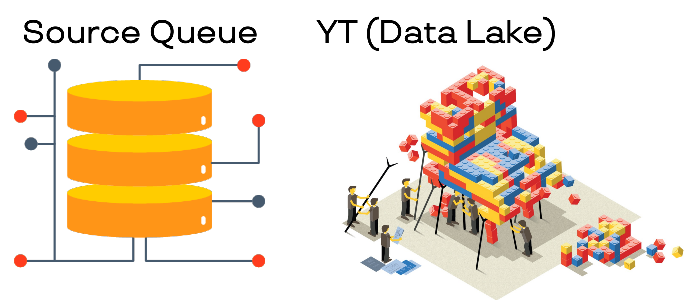


Архитектура слоев данных

Собрать	Стандартизировать	Сохранить	Предоставить	Проанализировать
RAW	ODS	DDS	CDM	REP

Source Queue

YT (Data Lake)



Greenplum (Data warehouse)

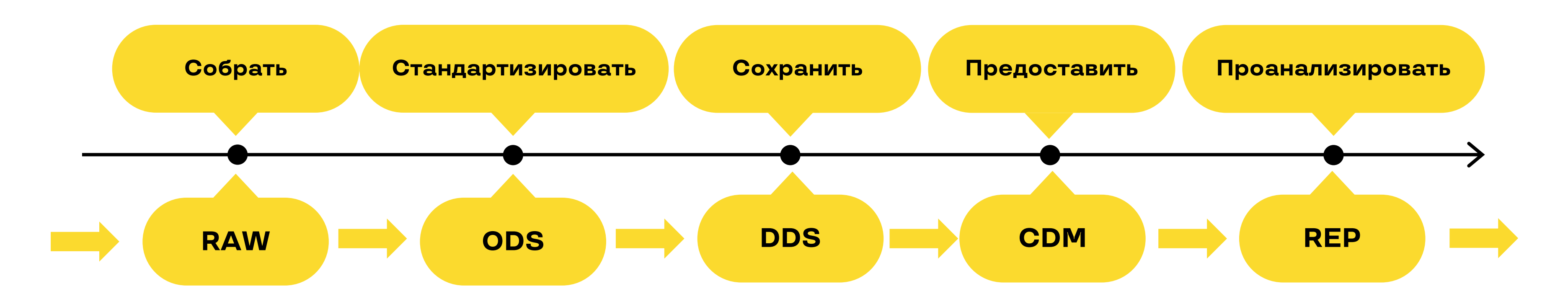


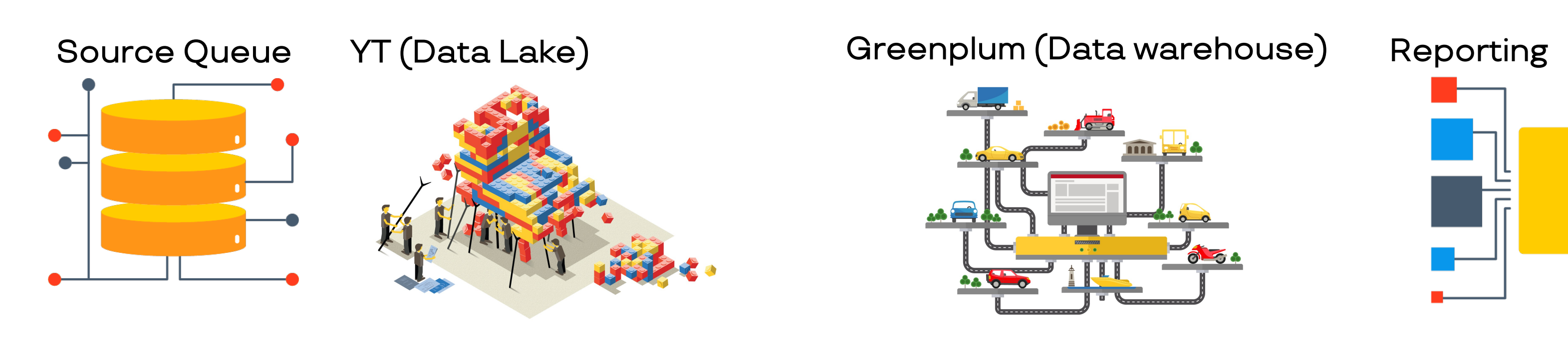
Reporting

- Кубы данных
- Отчеты и дашборды



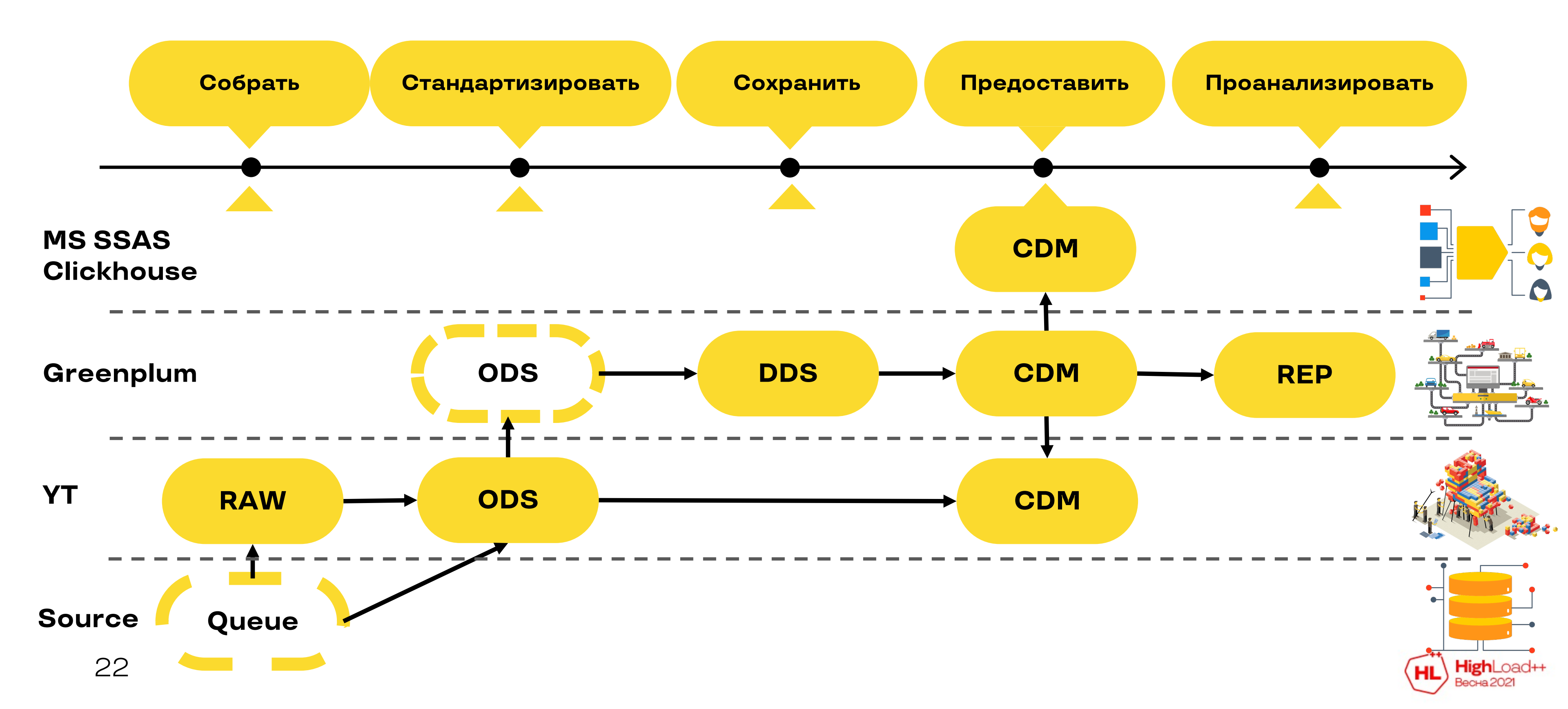
Архитектура слоев данных



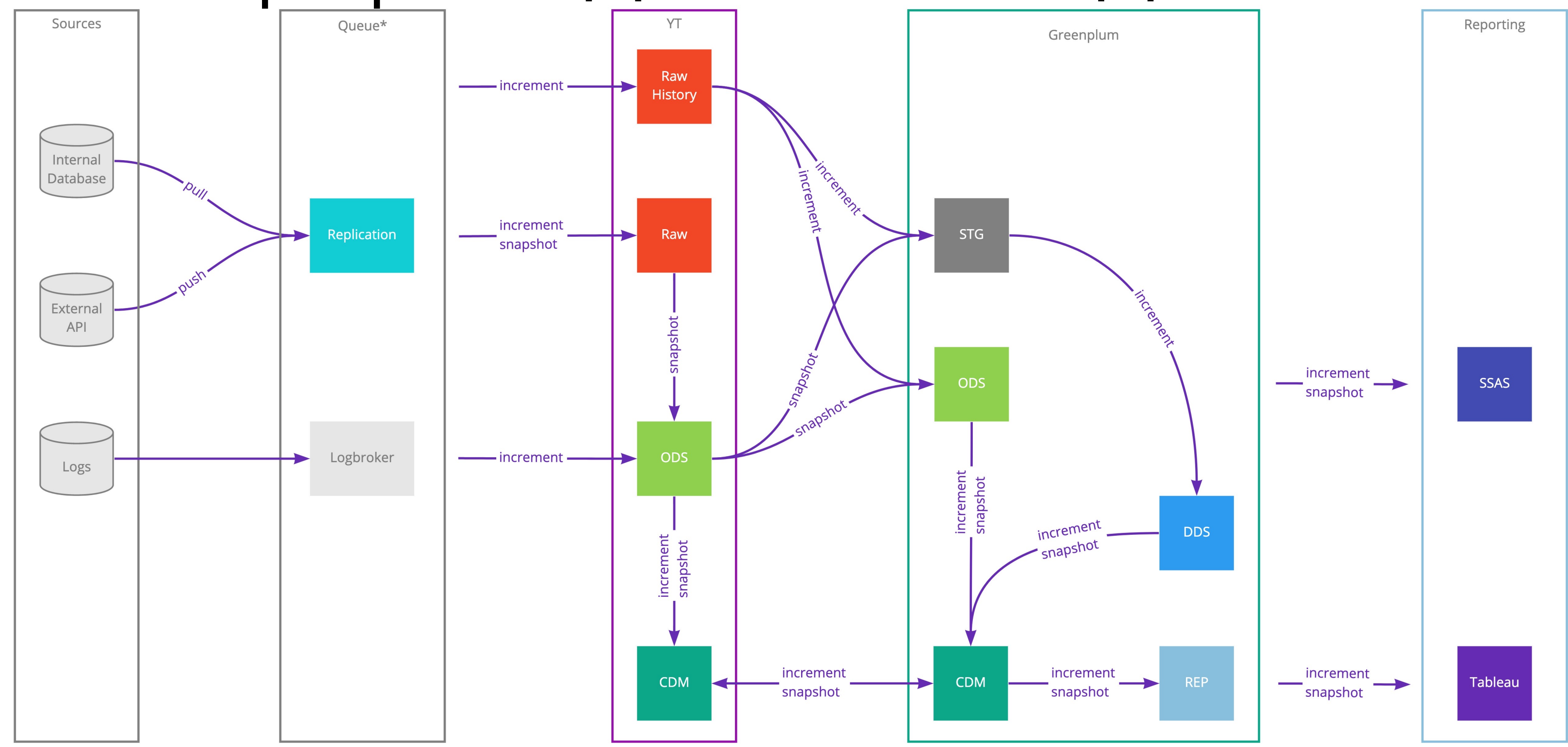




Архитектура слоев данных



Платформа данных Яндекс GO



 Владимир Верстов, https://smartdataconf.ru/2020/spb/talks/3fnm49atnbkyknqemzbtmc/

I. Зачем DWH метрики?

КАК РАЗВИВАТЬ?



Организация развития

Задачи платформы данных



Организация развития

Продуктовые

- Построение витрин
- Добавление новых источников
- Оптимизация ETL
- Развитие домена

Задачи платформы данных

Инфраструктурные

- Внедрение новых инструментов
- Повышение качества работы с данными



Организация развития

Скорость внедрения инициатив высокая

Долгосрочные проекты

+ ⟷ −

Продуктовые

- Построение витрин
- Добавление новых источников
- Оптимизация ETL
- Развитие домена

Задачи платформы данных

Инфраструктурные

- Внедрение новых инструментов
- Повышение качества работы с данными

− ⟷ +

Быстрый результат может быть важнее качества (в точке)

Качество архитектуры важнее сиюминутных задач



Домены данных

Данные сгруппированы по предметной области – домену (Domain)

- В одном домене может быть несколько объектов (таблиц)
- За несколько доменов отвечает одна команда
- Домены могут быть разных типов



Домены данных

Данные сгруппированы по предметной области – домену (Domain)

- В одном домене может быть несколько объектов (таблиц)
- За несколько доменов отвечает одна команда
- Домены могут быть разных типов

Source Domain

- Связаны с источником данных
- Структура подогнана под источник
- Включают в себя очистку, дедубликацию, приведение к стандартам и т.п.

Core Domain

- Связаны с крупной областью бизнеса
- Структура подогнана под минимизацию изменений
- Включает объединение данных из разных источников, генерацию суррогатных ключей и т.п.

Business Domain

- Связаны с потребителями данных
- Структура подогнана под удобства использования
- Фактически представляет собой специализированные витрины и/или отчеты



Домены данных

Собрать → RAW
Стандартизировать → ODS
Сохранить → DDS
Предоставить → CDM
Проанализировать → REP

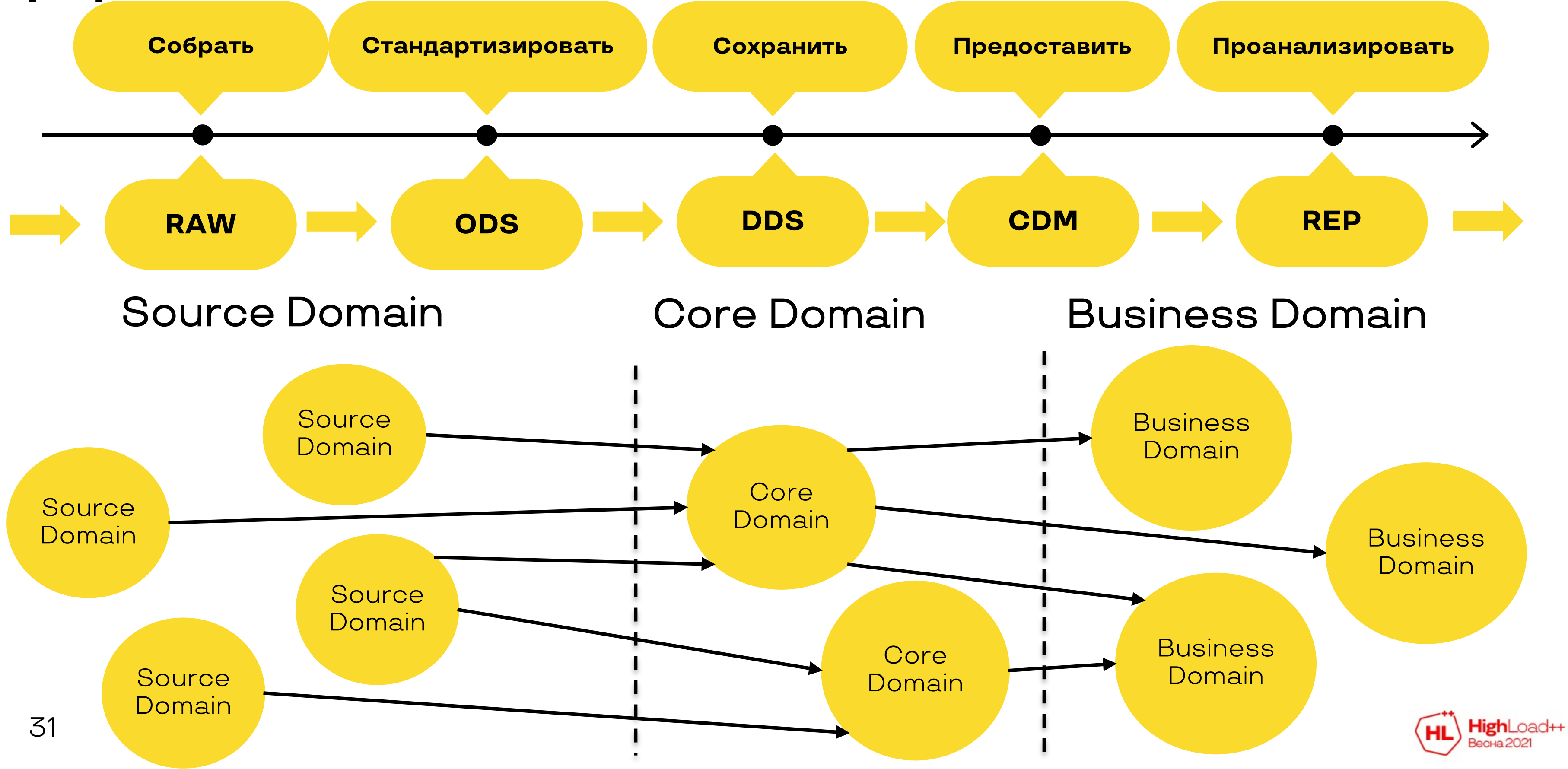
Source Domain

Core Domain

Business Domain



Домены данных

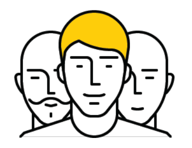


I. Зачем DWH метрики?

КАК УПРАВЛЯТЬ (БЕС)ПОРЯДКОМ?



Почему (бес)порядок?



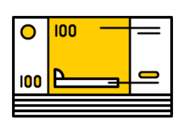
>500

уникальных пользователей данных в месяц



>900

отчетов по различным тематикам



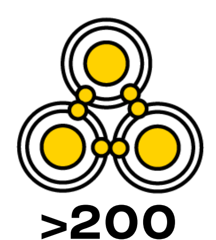
4

крупных бизнес-юнита: Такси, Еда, Лавка, Драйв



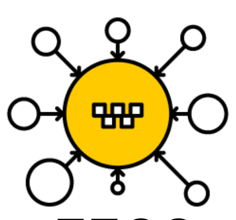
3 Пб

накопленных данных по четырем бизнес-юнитам



>200

доменов данных



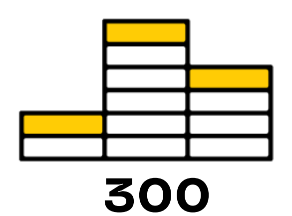
>3500

объектов хранилища



200

коммитов в день



300

merged Pull Request в месяц



I. Проблема:

развитием крупного DWH сложно управлять



Data Mesh

Global Governance & Open Standards
(enable interoperability)

Domain data owned & served by cross-functional team

New Data Domains can be created correlating data from other domains

Domain oriented data served as a product for any other domain to use

Domain's internal data pipeline uses common self-serve data infra

Data infra engineers

Data Infra as a Platform
(storage, pipeline, catalogue, Access control, etc.)

 Zhamak Dehghani, https://martinfowler.com/articles/data-monolith-to-mesh.html



Data Mesh

Global Governance & Open Standards
(enable interoperability)

Domain data owned & served by cross-functional team

New Data Domains can be created correlating data from other domains

Domain oriented data served as a product for any other domain to use

Domain's internal data pipeline uses common self-serve data infra

Data infra engineers

Data Infra as a Platform
(storage, pipeline, catalogue, Access control, etc.)

Продуктовые команды

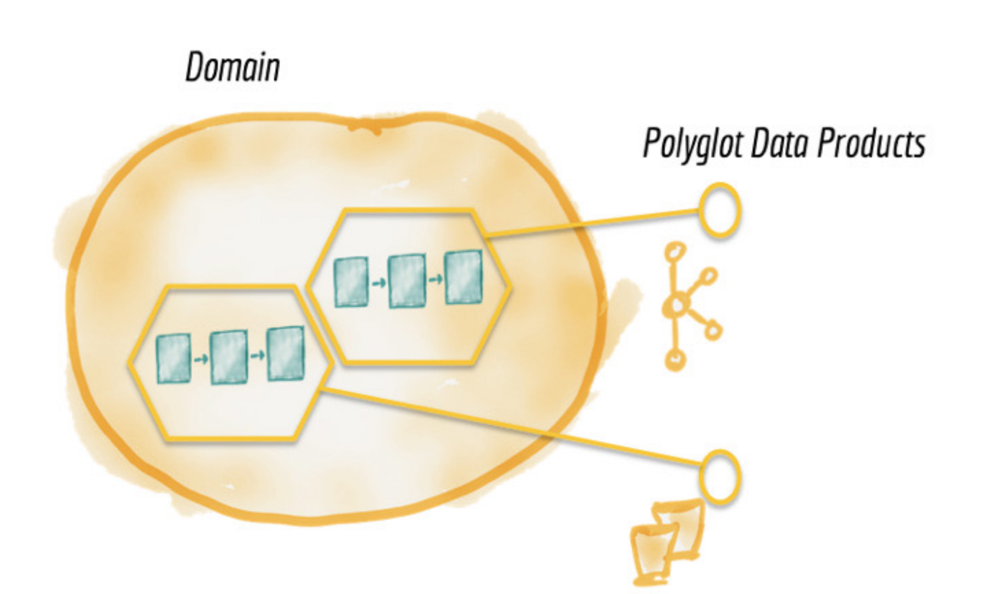
Домены данных

Инфраструктурные команды

 Zhamak Dehghani, https://martinfowler.com/articles/data-monolith-to-mesh.html



Данные – это продукт







ADDRESSABLE



TRUSTWORTHY
(DEFINED & MONITORED SLOs)



SELF-DESCRIBING



INTER OPERABLE
(GOVERENED BY OPEN STANDARDS)



SECURE



 Zhamak Dehghani, https://martinfowler.com/articles/data-monolith-to-mesh.html



I. Проблема:

развитием крупного DWH сложно управлять

II. Решение:

покрыть работу DWH метриками



Данные – это продукт

Витрины, измерения, любые наборы данных – это продукт

Аналитики, DS, ML-специалисты, менеджеры – пользователи продукта

- Удобство использования
- Частота использования
- Легкость обнаружения
- Качество данных
- Понятное описание семантики
- Интегрируемость данных и стандарты



Данные – это продукт

Витрины, измерения, любые наборы данных – это продукт

Аналитики, DS, ML-специалисты, менеджеры – пользователи продукта

- Удобство использования
- Частота использования
- Легкость обнаружения
- Качество данных
- Понятное описание семантики
- Интегрируемость данных и стандарты

Покроем метриками



Данные – это продукт

Продуктовая команда – независимая единица поставки счастья

Data Partner



Владелец данных (= продукта)

- Коммуникации с пользователем
- Управление требованиями
- Развитие домена
- Постановка задач
- Создание метаданных витрин/отчетов

Data Engineer



Разработчик данных (= продукта)

- Выполнение задач на разработку
- Реализация ETL/ELT на базе платформы
- Создание сложных алгоритмов агрегации данных/подсчетов
- Физическая реализация метаданных на доступных инструментах



Данные – это продукт

Продуктовая команда – независимая единица поставки счастья

Data Partner



Владелец данных (= продукта)

- Коммуникации с пользователем
- Управление требованиями
- Развитие домена
- Постановка задач
- Создание метаданных витрин/отчетов

Data Engineer



Разработчик данных (= продукта)

- Выполнение задач на разработку
- Реализация ETL/ELT на базе платформы
- Создание сложных алгоритмов агрегации данных/подсчетов
- Физическая реализация метаданных на доступных инструментах

Работу продуктовых команд будем оценивать через метрики



I. ЗАЧЕМ DWH МЕТРИКИ?

II. КАК РЕАЛИЗОВАЛИ?

III. ЧТО ПОЛУЧИЛИ?

IV. СТОИЛО ЛИ ТОГО?



I. Проблема:

развитием крупного DWH сложно управлять

II. Решение:

покрыть работу DWH метриками

III. Идея:

использовать данные систем DWH в самом DWH («DWH для DWH»)



DHW для DWH

Почему бы не рассмотреть DWH как источник информации для самого DWH?

Транзакционная информация

Что происходит?

- Логи обращения к Greenplum
- Логи обращения к YT
- Логи обращения к Tableau
- Логи обращения к MS SSAS
- Логи ошибок по объектам
- Информация об отставании данных

Статическая информация

С чем происходит?

- Метаданные систем
- Данные из системы учета пользователей
- Метаданные из нашего репозитория метаданных
- Граф связей между тасками ETL-процессов



MetaDWH

Source Domain

Greenplum
- Логи использования
- Метаданные объектов

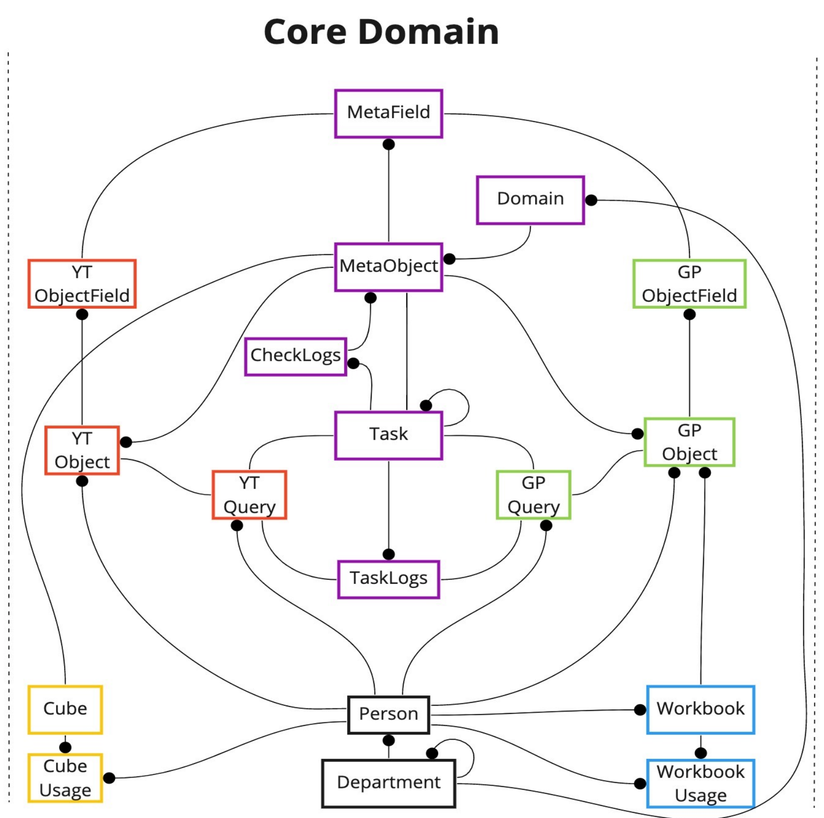
YT
- Логи использования
- Метаданные объектов

Tableau
- Логи использования
- Метаданные объектов

MS SSAS
- Логи использования
- Метаданные объектов

Staff
- Профиль пользователя
- Орг.Структура

Platform
- Метаданные тасок
- Логи запусков
- Логи проверок
- Метаданные объектов



Business Domain

Техническая информация
- Витрина по размеру данных
- Витрина по потреблению ресурсов

Использование объектов
- Витрина по использованию объектов
- Витрина по использованию отчетов

Витрины с метаданными
- Метаданные объектов
- Витрина по таскам
- Витрина по результатам проверок



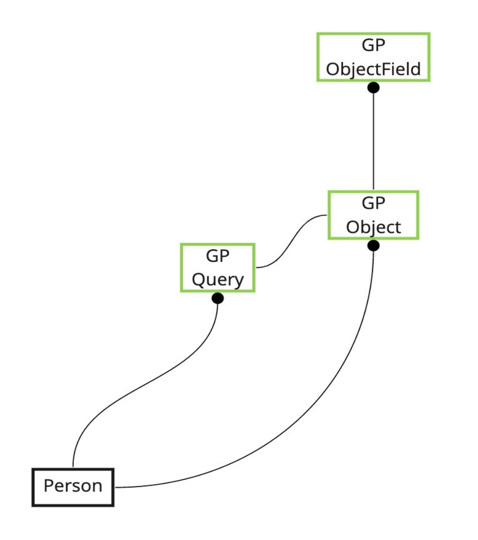
Greenplum

Source Domain | **Core Domain** | **Business Domain**

Greenplum

Логи использования

Метаданные объектов





Staff

Source Domain

Greenplum

Логи использования

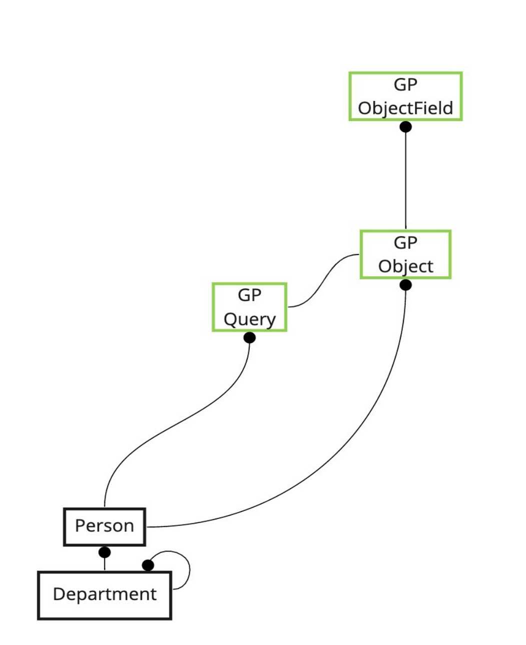
Метаданные объектов

Staff

Профиль пользователя

Орг.Структура

Core Domain



Business Domain



YT

Source Domain



Core Domain

Business Domain

Greenplum

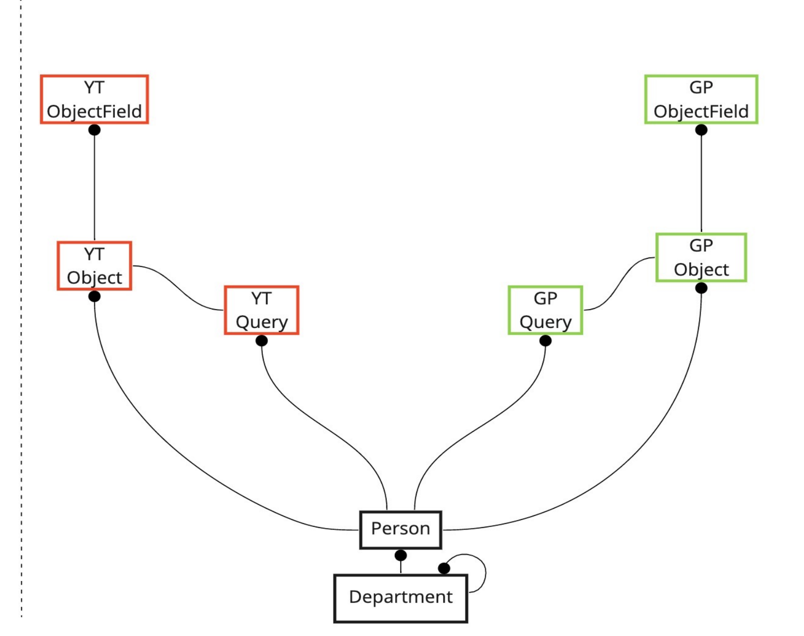
Логи использования

Метаданные объектов

YT

Логи использования

Метаданные объектов



Staff

Профиль пользователя

Орг.Структура

MS SSAS

Source Domain

Greenplum

Логи использования

Метаданные объектов

YT

Логи использования

Метаданные объектов

MS SSAS

Логи использования

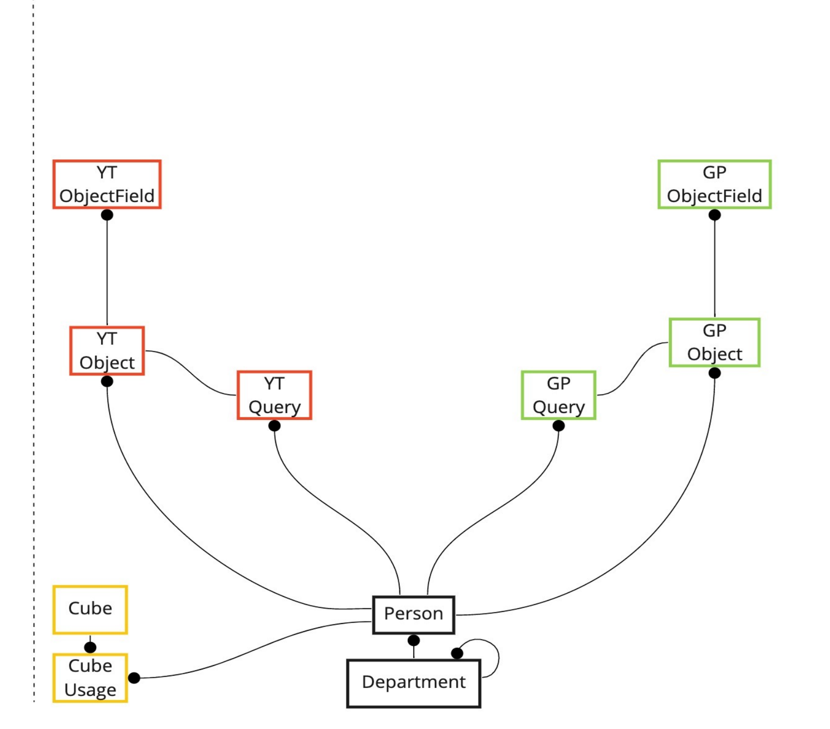
Метаданные объектов

Staff

Профиль пользователя

Орг.Структура

Core Domain



Business Domain



Потребление ресурсов

Source Domain

Greenplum

Логи использования

Метаданные объектов

YT

Логи использования

Метаданные объектов

MS SSAS

Логи использования

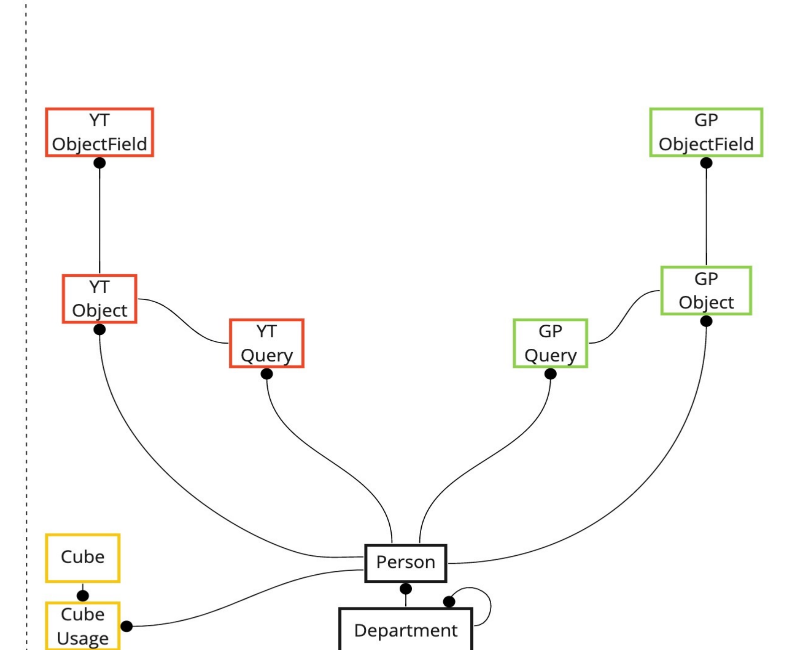
Метаданные объектов

Staff

Профиль пользователя

Орг.Структура

Core Domain



Business Domain

Техническая информация

Витрина по размеру данных

Витрина по потреблению ресурсов



Tableau

Source Domain

Greenplum
Логи использования

Метаданные объектов

YT
Логи использования

Метаданные объектов

Tableau
Логи использования

Метаданные объектов

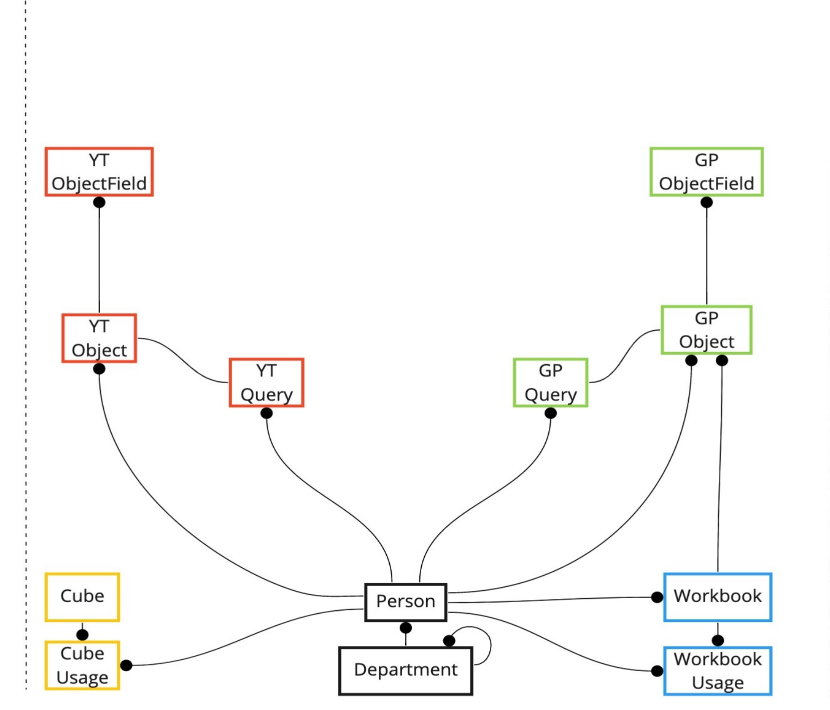
MS SSAS
Логи использования

Метаданные объектов

Staff
Профиль пользователя

Орг.Структура

Core Domain



Business Domain

Техническая информация

Витрина по размеру данных

Витрина по потреблению ресурсов



Использование объектов

Source Domain

Greenplum
- Логи использования
- Метаданные объектов

YT
- Логи использования
- Метаданные объектов

Tableau
- Логи использования
- Метаданные объектов

MS SSAS
- Логи использования
- Метаданные объектов

Staff
- Профиль пользователя
- Орг.Структура

Core Domain

Business Domain

Техническая информация
- Витрина по размеру данных
- Витрина по потреблению ресурсов

Использование объектов
- Витрина по использованию объектов
- Витрина по использованию отчетов



Метаданные объектов

Source Domain

Greenplum
- Логи использования
- Метаданные объектов

YT
- Логи использования
- Метаданные объектов

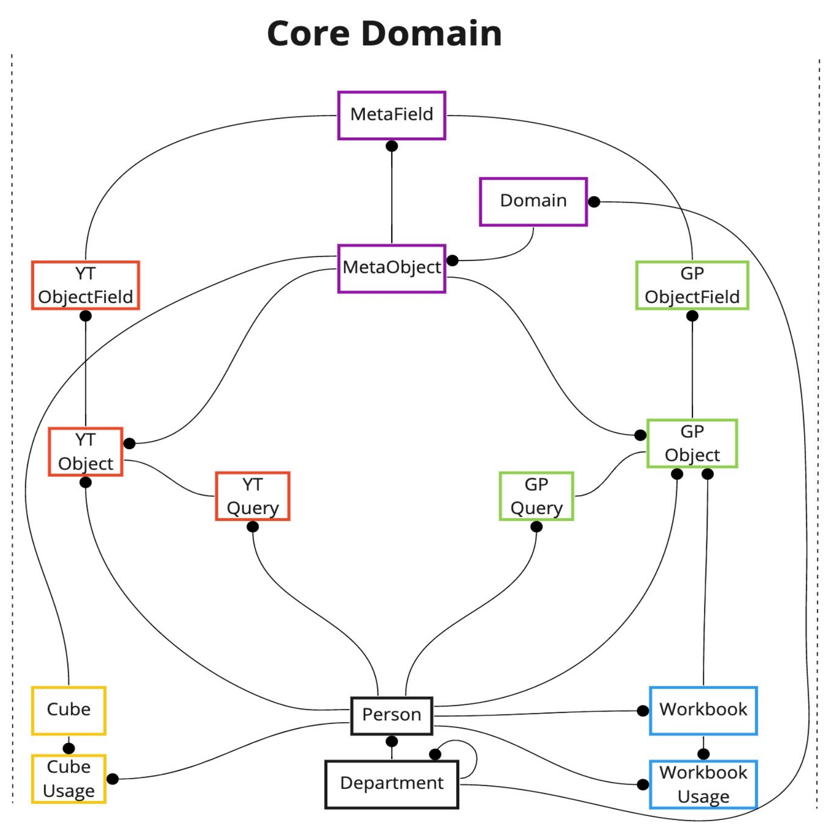
Tableau
- Логи использования
- Метаданные объектов

MS SSAS
- Логи использования
- Метаданные объектов

Staff
- Профиль пользователя
- Орг.Структура

Platform
- Метаданные объектов

Core Domain



Business Domain

Техническая информация
- Витрина по размеру данных
- Витрина по потреблению ресурсов

Использование объектов
- Витрина по использованию объектов
- Витрина по использованию отчетов



Витрины по метаданным

Source Domain

Greenplum
- Логи использования
- Метаданные объектов

YT
- Логи использования
- Метаданные объектов

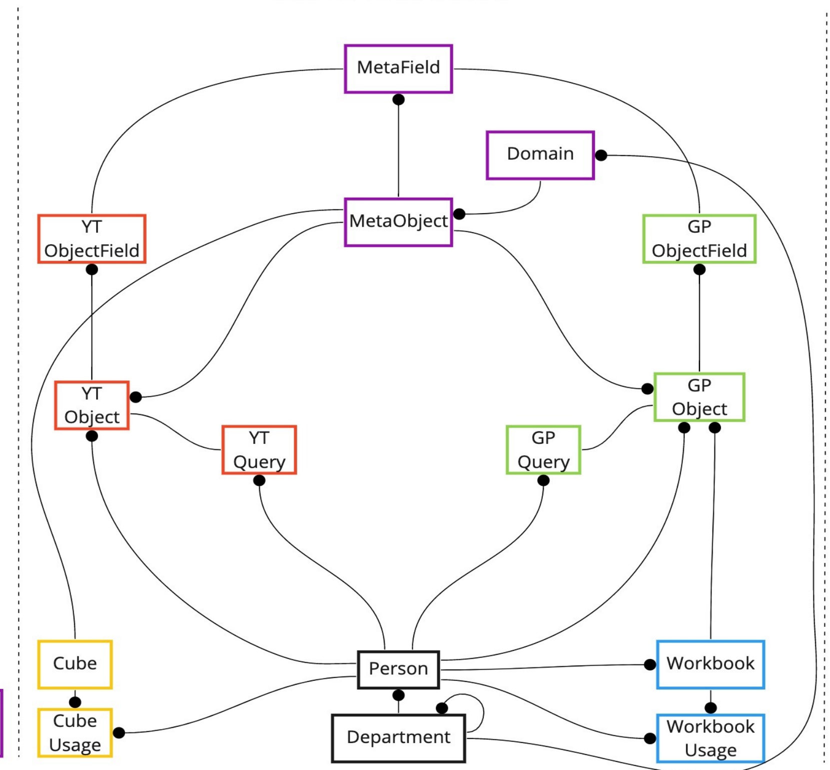
Tableau
- Логи использования
- Метаданные объектов

MS SSAS
- Логи использования
- Метаданные объектов

Staff
- Профиль пользователя
- Орг.Структура

Platform
- Метаданные объектов

Core Domain



Business Domain

Техническая информация
- Витрина по размеру данных
- Витрина по потреблению ресурсов

Использование объектов
- Витрина по использованию объектов
- Витрина по использованию отчетов

Витрины с метаданными
- Метаданные объектов



Знания о запусках ETL

Source Domain

Greenplum
- Логи использования
- Метаданные объектов

YT
- Логи использования
- Метаданные объектов

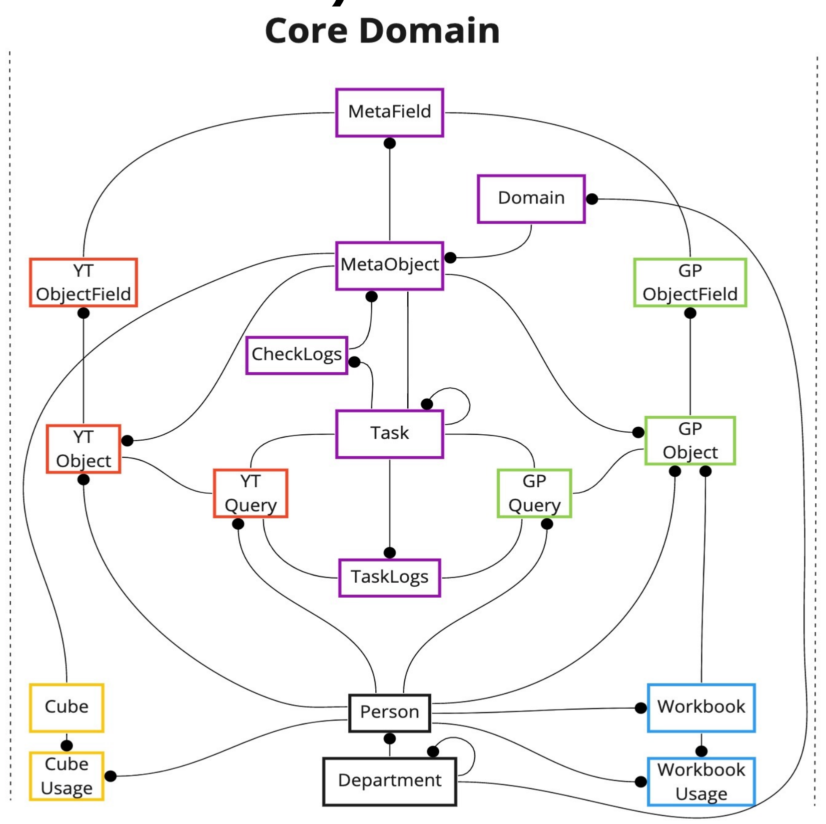
Tableau
- Логи использования
- Метаданные объектов

MS SSAS
- Логи использования
- Метаданные объектов

Staff
- Профиль пользователя
- Орг.Структура

Platform
- Метаданные тасок
- Логи запусков
- Логи проверок
- Метаданные объектов

Core Domain



Business Domain

Техническая информация
- Витрина по размеру данных
- Витрина по потреблению ресурсов

Использование объектов
- Витрина по использованию объектов
- Витрина по использованию отчетов

Витрины с метаданными
- Метаданные объектов



MetaDWH

Source Domain

Greenplum
- Логи использования
- Метаданные объектов

YT
- Логи использования
- Метаданные объектов

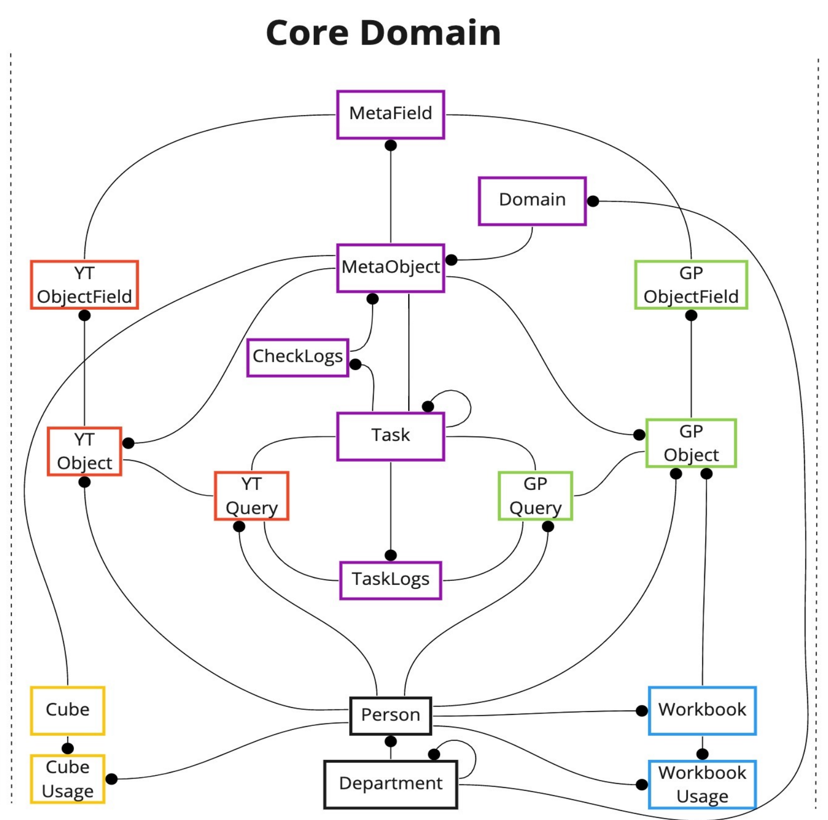
Tableau
- Логи использования
- Метаданные объектов

MS SSAS
- Логи использования
- Метаданные объектов

Staff
- Профиль пользователя
- Орг.Структура

Platform
- Метаданные тасок
- Логи запусков
- Логи проверок
- Метаданные объектов

Core Domain



Business Domain

Техническая информация
- Витрина по размеру данных
- Витрина по потреблению ресурсов

Использование объектов
- Витрина по использованию объектов
- Витрина по использованию отчетов

Витрины с метаданными
- Метаданные объектов
- Витрина по таскам
- Витрина по результатам проверок



I. ЗАЧЕМ DWH МЕТРИКИ?

II. КАК РЕАЛИЗОВАЛИ?

III. ЧТО ПОЛУЧИЛИ?

IV. СТОИЛО ЛИ ТОГО?



I. Проблема:

развитием крупного DWH сложно управлять

II. Решение:

покрыть работу DWH метриками

III. Идея:

использовать данные систем DWH в самом DWH («DWH для DWH»)

IV. Результат:

аналитика по работе и развитию самого DWH



III. Что получили?

ЧЕМ ПОЛЬЗУЮТСЯ В ХРАНИЛИЩЕ?



Профиль объекта

Object profile

Datasource updated at 15.09.2020 12:26:12

Object usage	Include Robots	DMP activity	Sort by	System	Object	
Last 3 months	True	False	Requests	YT	eda_etl_new / ...	Apply

General info

Object	dm_order
Path	//home/eda-dwh/cdm/order/dm_order
System	yt
Layer	cdm
Domain	[Business Domain] eda_etl_new.order
DMP Team	
Production flag	✓ True
CDM or REP flag	✓ True
Deprecated flag	× False
Tech flag	× False

6 077 **Usage** (163 … 263)

35 **Users** (5 … 8)

Daily usage dynamic

00 05 10 15 20

Department

Name

	Usage	Users	Daily Usage
[illegible]	1 344	1	
[illegible]	878	3	
[illegible]	850	1	
[illegible]	728	1	
[illegible]	677	5	
[illegible]	555	1	
[illegible]	514	2	
[illegible]	270	3	
[illegible]	86	1	
[illegible]	55	1	
[illegible]	50	1	
[illegible]	29	1	

User

Staff login	Usage		Daily Usage
[illegible]	1 344	22,1%	
[illegible]	872	14,3%	
[illegible]	850	14,0%	
[illegible]	728	12,0%	
[illegible]	555	9,1%	
[illegible]	475	7,8%	
[illegible]	376	6,2%	
[illegible]	192	3,2%	
[illegible]	100	1,6%	
[illegible]	87	1,4%	
[illegible]	86	1,4%	
[illegible]	83	1,4%	

Fields

Name	Type	Description
cancel_reason_group_..	String	Название группы
cancel_reason_id	Int	Идентификатор прич..
cancel_reason_name	String	Описание причины от..
cancel_reasons_syste..	Boolean	Флаг автоматической..
cancelled_lat	Double	Широта координат ку..
cancelled_lon	Double	Долгота координат ку..
commission_value_w_v..	Double	Сумма комиссии, кот..
confirmed_flg	Boolean	Индикатор подтверж..
cooking_type	String	Тип готовки. *Опреде..
corp_order_flg	Boolean	Индикатор корпорати..
country_id	Int	Идентификатор стра..
country_name	String	Наименование страны
courier_assigned_lat	Double	Широта координат ку..
courier_assigned_lon	Double	Долгота координат ку..
courier_balance_id	Int	Идентификатор курь..
courier_delay_sec	Int	Опоздание курьера в..
courier_id	Int	Идентификатор курь..
courier_selfemploeyed_..	Boolean	Индикатор самозанят..
courier_service_id	Int	Идентификатор курь..
courier_service_income..	Int	Идентификатор дохо..
courier_service_name	String	Наименование курье..
courier_type	String	Тип курьера (пеший, ..
courier_type_code	String	Тип передвижения ку..
courier_username	String	Фамилия и имя курье..



Профиль объекта

Object profile

Datasource updated at 15.09.2020 12:26:12

Object usage	Include Robots	DMP activity	Sort by	System	Object	
Last 3 months	True	False	Requests	YT	eda_etl_new / ...	Apply

General info

Object	dm_order
Path	//home/eda-dwh/cdm/order/dm_order
System	yt
Layer	cdm
Domain	[Business Domain] eda_etl_new.order
DMP Team	
Production flag	✓ True
CDM or REP flag	✓ True
Deprecated flag	× False
Tech flag	× False

6 077
Usage
163 263

35
Users
5 8

Daily usage dynamic

00 05 10 15 20

Department

Name

	Usage	Users	Daily Usage
[illegible]	1 344	1	
[illegible]	878	3	
[illegible]	850	1	
[illegible]	728	1	
[illegible]	677	5	
[illegible]	555	1	
[illegible]	514	2	
[illegible]	270	3	
[illegible]	86	1	
[illegible]	55	1	
[illegible]	50	1	
[illegible]	29	1	

User

Staff login	Usage		Daily Usage
[illegible]	1 344	22,1%	
[illegible]	872	14,3%	
[illegible]	850	14,0%	
[illegible]	728	12,0%	
[illegible]	555	9,1%	
[illegible]	475	7,8%	
[illegible]	376	6,2%	
[illegible]	192	3,2%	
[illegible]	100	1,6%	
[illegible]	87	1,4%	
[illegible]	86	1,4%	
[illegible]	83	1,4%	

Fields

Name	Type	Description
cancel_reason_group_..	String	Название группы
cancel_reason_id	Int	Идентификатор прич..
cancel_reason_name	String	Описание причины от..
cancel_reasons_syste..	Boolean	Флаг автоматической..
cancelled_lat	Double	Широта координат ку..
cancelled_lon	Double	Долгота координат ку..
commission_value_w_v..	Double	Сумма комиссии, кот..
confirmed_flg	Boolean	Индикатор подтверж..
cooking_type	String	Тип готовки. *Опреде..
corp_order_flg	Boolean	Индикатор корпорати..
country_id	Int	Идентификатор стра..
country_name	String	Наименование страны
courier_assigned_lat	Double	Широта координат ку..
courier_assigned_lon	Double	Долгота координат ку..
courier_balance_id	Int	Идентификатор курь..
courier_delay_sec	Int	Опоздание курьера в..
courier_id	Int	Идентификатор курь..
courier_selfemploeyed_..	Boolean	Индикатор самозанят..
courier_service_id	Int	Идентификатор курь..
courier_service_income..	Int	Идентификатор дохо..
courier_service_name	String	Наименование курье..
courier_type	String	Тип курьера (пеший, ..
courier_type_code	String	Тип передвижения ку..
courier_username	String	Фамилия и имя курье..



Профиль объекта

Object profile

Datasource updated at 15.09.2020 12:26:12

Object usage	Include Robots	DMP activity	Sort by	System	Object	
Last 3 months	True	False	Requests	YT	eda_etl_new / ...	Apply

General info

Object	dm_order
Path	//home/eda-dwh/cdm/order/dm_order
System	yt
Layer	cdm
Domain	[Business Domain] eda_etl_new.order
DMP Team	
Production flag	✓ True
CDM or REP flag	✓ True
Deprecated flag	× False
Tech flag	× False

6 077
Usage

163 263

35
Users

5 8

Daily usage dynamic

00 05 10 15 20

Department

Name

	Usage	Users	Daily Usage
[illegible]	1 344	1	
[illegible]	878	3	
[illegible]	850	1	
[illegible]	728	1	
[illegible]	677	5	
[illegible]	555	1	
[illegible]	514	2	
[illegible]	270	3	
[illegible]	86	1	
[illegible]	55	1	
[illegible]	50	1	
[illegible]	29	1	

User

Staff login	Usage		Daily Usage
[illegible]	1 344	22,1%	
[illegible]	872	14,3%	
[illegible]	850	14,0%	
[illegible]	728	12,0%	
[illegible]	555	9,1%	
[illegible]	475	7,8%	
[illegible]	376	6,2%	
[illegible]	192	3,2%	
[illegible]	100	1,6%	
[illegible]	87	1,4%	
[illegible]	86	1,4%	
[illegible]	83	1,4%	

Fields

Name	Type	Description
cancel_reason_group_..	String	Название группы
cancel_reason_id	Int	Идентификатор прич..
cancel_reason_name	String	Описание причины от..
cancel_reasons_syste..	Boolean	Флаг автоматической..
cancelled_lat	Double	Широта координат ку..
cancelled_lon	Double	Долгота координат ку..
commission_value_w_v..	Double	Сумма комиссии, кот..
confirmed_flg	Boolean	Индикатор подтверж..
cooking_type	String	Тип готовки. *Опреде..
corp_order_flg	Boolean	Индикатор корпорати..
country_id	Int	Идентификатор стра..
country_name	String	Наименование страны
courier_assigned_lat	Double	Широта координат ку..
courier_assigned_lon	Double	Долгота координат ку..
courier_balance_id	Int	Идентификатор курь..
courier_delay_sec	Int	Опоздание курьера в..
courier_id	Int	Идентификатор курь..
courier_selfemploeyed_..	Boolean	Индикатор самозанят..
courier_service_id	Int	Идентификатор курь..
courier_service_income..	Int	Идентификатор дохо..
courier_service_name	String	Наименование курье..
courier_type	String	Тип курьера (пеший, ..
courier_type_code	String	Тип передвижения ку..
courier_username	String	Фамилия и имя курье..



Профиль объекта

Object profile

Datasource updated at 15.09.2020 12:26:12

Object usage	Include Robots	DMP activity	Sort by	System	Object	
Last 3 months	True	False	Requests	YT	eda_etl_new / ...	Apply

General info

Object	dm_order
Path	//home/eda-dwh/cdm/order/dm_order
System	yt
Layer	cdm
Domain	[Business Domain] eda_etl_new.order
DMP Team	
Production flag	✓ True
CDM or REP flag	✓ True
Deprecated flag	× False
Tech flag	× False

6 077
Usage
163 263

35
Users
5 8

Daily usage dynamic

00 05 10 15 20

Department

Name

	Usage	Users	Daily Usage
[illegible]	1 344	1	
[illegible]	878	3	
[illegible]	850	1	
[illegible]	728	1	
[illegible]	677	5	
[illegible]	555	1	
[illegible]	514	2	
[illegible]	270	3	
[illegible]	86	1	
[illegible]	55	1	
[illegible]	50	1	
[illegible]	29	1	

User

Staff login	Usage		Daily Usage
[illegible]	1 344	22,1%	
[illegible]	872	14,3%	
[illegible]	850	14,0%	
[illegible]	728	12,0%	
[illegible]	555	9,1%	
[illegible]	475	7,8%	
[illegible]	376	6,2%	
[illegible]	192	3,2%	
[illegible]	100	1,6%	
[illegible]	87	1,4%	
[illegible]	86	1,4%	
[illegible]	83	1,4%	

Fields

Name	Type	Description
cancel_reason_group_..	String	Название группы
cancel_reason_id	Int	Идентификатор прич..
cancel_reason_name	String	Описание причины от..
cancel_reasons_syste..	Boolean	Флаг автоматической..
cancelled_lat	Double	Широта координат ку..
cancelled_lon	Double	Долгота координат ку..
commission_value_w_v..	Double	Сумма комиссии, кот..
confirmed_flg	Boolean	Индикатор подтверж..
cooking_type	String	Тип готовки. *Опреде..
corp_order_flg	Boolean	Индикатор корпорати..
country_id	Int	Идентификатор стра..
country_name	String	Наименование страны
courier_assigned_lat	Double	Широта координат ку..
courier_assigned_lon	Double	Долгота координат ку..
courier_balance_id	Int	Идентификатор курь..
courier_delay_sec	Int	Опоздание курьера в..
courier_id	Int	Идентификатор курь..
courier_selfemploeyed_..	Boolean	Индикатор самозанят..
courier_service_id	Int	Идентификатор курь..
courier_service_income..	Int	Идентификатор дохо..
courier_service_name	String	Наименование курье..
courier_type	String	Тип курьера (пеший, ..
courier_type_code	String	Тип передвижения ку..
courier_username	String	Фамилия и имя курье..



Использование объектов

Object usage dashboard

Datasource updated at 15.09.2020 8:16:42

Report period	Report date	Chart type	DAU/CNT	System	Robots	DMP activity	Production flg
Last 4 months	13.09.2020	Treemap	DAU	(All)	False	False	True

Object

Select Object to drill down

order 1,4%
order 1,3%
place 1,5%
user 1,0%
order_proc 1,8%
orders 0,7%
order 1,3%
order 0,6%
tickets
city
park
dm_order 8,2%
dim_driver 2,2%
dm_order 2,4%

Department

Группа аналитики новых продуктов 4,6%
Группа
Служба
Группа
Группа
Служба
Группа
Группа
Команда

General info

14.09.2020

485 — 709 Used Objects

4,2% — 22,6% Prod Usage %

329 — 344 Users

7 057 — 9 157 Usage DAU

System

GP	11,5%
YT	88,5%

Schema

Production

	Production
cdm	21,89%
dds	8,89%
ods	39,25%
raw	25,60%
rep	0,00%
tech	4,29%
udm	0,08%

Team

Select Team to drill down

Atlas	0,01%
Demand	10,20%
Marketplace	30,01%
Supply Eda	34,71%
Supply Taxi	22,65%

User

	Usage	Objects
	4 020	109
	3 426	106
	2 847	102
	2 774	128
	2 564	56
	2 549	85
	2 303	97
	2 109	42
	2 013	76
	1 896	88
	1 853	125
	1 843	122
	1 835	72
	1 736	115
	1 707	65
	1 630	168

Domain

	Usage	Objects
[Source Domain] eda_etl_new.bigfood	37 770	365
[Unknown Domain]	28 682	117
[Source Domain] taxi_etl.mdb	19 138	178
[Source Domain] taxi_etl.taximeter	7 573	61
[Business Domain] taxi_etl.geo	3 723	25
[Source Domain] taxi_etl.appmetrica	3 600	29
[Source Domain] eda_etl.bigfood	3 420	88
[Source Domain] taxi_etl.da	2 132	8
[Business Domain] eda_etl.marketplace	2 090	4
[Business Domain] taxi_etl.supply	1 969	44
[Source Domain] taxi_etl.zendesk_hiring	1 716	9
[Source Domain] taxi_etl.googledocs	1 245	21
[Source Domain] eda_etl_new.lavka_1c	1 228	31
[Business Domain] core_etl.geo	1 173	9
[Source Domain] taxi_etl.amocrm	1 106	11



Использование объектов

Object usage dashboard
Datasource updated at 15.09.2020 8:16:42

Report period	Report date	Chart type	DAU/CNT	System	Robots	DMP activity	Production flg
Last 4 months	13.09.2020	Treemap	DAU	(All)	False	False	True

Object
Select Object to drill down

order 1,4%
order 1,3%
place 1,5%
user 1,0%
order_proc 1,8%
orders 0,7%
order 1,3%
order 0,6%
tickets
city
park
dm_order 8,2%
dim_driver 2,2%
dm_order 2,4%

Department

Группа аналитики новых продуктов 4,6%
Группа
Служба
Группа
Группа
Служба
Группа
Группа
Команда

General info 14.09.2020

485 — 709 Used Objects
4,2% — 22,6% Prod Usage %
329 — 344 Users
7 057 — 9 157 Usage DAU

System

GP	11,5%
YT	88,5%

Schema Production

Production

cdm	21,89%
dds	8,89%
ods	39,25%
raw	25,60%
rep	0,00%
tech	4,29%
udm	0,08%

Team
Select Team to drill down

Atlas	0,01%
Demand	10,20%
Marketplace	30,01%
Supply Eda	34,71%
Supply Taxi	22,65%

User

Usage	Objects
4 020	109
3 426	106
2 847	102
2 774	128
2 564	56
2 549	85
2 303	97
2 109	42
2 013	76
1 896	88
1 853	125
1 843	122
1 835	72
1 736	115
1 707	65
1 630	168

Domain

	Usage	Objects
[Source Domain] eda_etl_new.bigfood	37 770	365
[Unknown Domain]	28 682	117
[Source Domain] taxi_etl.mdb	19 138	178
[Source Domain] taxi_etl.taximeter	7 573	61
[Business Domain] taxi_etl.geo	3 723	25
[Source Domain] taxi_etl.appmetrica	3 600	29
[Source Domain] eda_etl.bigfood	3 420	88
[Source Domain] taxi_etl.da	2 132	8
[Business Domain] eda_etl.marketplace	2 090	4
[Business Domain] taxi_etl.supply	1 969	44
[Source Domain] taxi_etl.zendesk_hiring	1 716	9
[Source Domain] taxi_etl.googledocs	1 245	21
[Source Domain] eda_etl_new.lavka_1c	1 228	31
[Business Domain] core_etl.geo	1 173	9
[Source Domain] taxi_etl.amocrm	1 106	11



Использование объектов

Object usage dashboard
Datasource updated at 15.09.2020 8:16:42

Report period	Report date	Chart type	DAU/CNT	System	Robots	DMP activity	Production flg
Last 4 months	13.09.2020	Treemap	DAU	(All)	False	False	True

Object
Select Object to drill down

order 1,4%
order 1,3%
place 1,5%
user 1,0%
order_proc 1,8%
orders 0,7%
order 1,3%
order 0,6%
tickets
city
park
dm_order 8,2%
dm_driver 2,2%
dm_order 2,4%

Department

Группа аналитики новых продуктов 4,6%
Группа
Служба
Группа
Группа
Служба
Группа
Группа
Команда

General info 14.09.2020

485
709 Used Objects

4,2%
22,6% Prod Usage %

329
344 Users

7 057
9 157 Usage DAU

System

GP	11,5%
YT	88,5%

Schema Production

Production

cdm	21,89%
dds	8,89%
ods	39,25%
raw	25,60%
rep	0,00%
tech	4,29%
udm	0,08%

Team
Select Team to drill down

Atlas	0,01%
Demand	10,20%
Marketplace	30,01%
Supply Eda	34,71%
Supply Taxi	22,65%

User

	Usage	Objects
	4 020	109
	3 426	106
	2 847	102
	2 774	128
	2 564	56
	2 549	85
	2 303	97
	2 109	42
	2 013	76
	1 896	88
	1 853	125
	1 843	122
	1 835	72
	1 736	115
	1 707	65
	1 630	168

Domain

	Usage	Objects
[Source Domain] eda_etl_new.bigfood	37 770	365
[Unknown Domain]	28 682	117
[Source Domain] taxi_etl.mdb	19 138	178
[Source Domain] taxi_etl.taximeter	7 573	61
[Business Domain] taxi_etl.geo	3 723	25
[Source Domain] taxi_etl.appmetrica	3 600	29
[Source Domain] eda_etl.bigfood	3 420	88
[Source Domain] taxi_etl.da	2 132	8
[Business Domain] eda_etl.marketplace	2 090	4
[Business Domain] taxi_etl.supply	1 969	44
[Source Domain] taxi_etl.zendesk_hiring	1 716	9
[Source Domain] taxi_etl.googledocs	1 245	21
[Source Domain] eda_etl_new.lavka_1c	1 228	31
[Business Domain] core_etl.geo	1 173	9
[Source Domain] taxi_etl.amocrm	1 106	11
[illegible]	[illegible]	[illegible]



Использование объектов

Object usage dashboard

Datasource updated at 15.09.2020 8:16:42

Report period	Report date	Chart type	DAU/CNT	System	Robots	DMP activity	Production flg
Last 4 months	13.09.2020	Treemap	DAU	(All)	False	False	True

Object

Select Object to drill down

order 1,4%
order 1,3%
place 1,5%
user 1,0%
order_proc 1,8%
orders 0,7%
order 1,3%
order 0,6%
tickets
city
park
dm_order 8,2%
dm_driver 2,2%
dm_order 2,4%

Department

Группа аналитики новых продуктов 4,6%
Группа
Служба
Группа
Группа
Служба
Группа
Группа
Команда

General info

14.09.2020

709 Used Objects (485)

22,6% Prod Usage % (4,2%)

344 Users (329)

9 157 Usage DAU (7 057)

System

System	Share
GP	11,5%
YT	88,5%

Schema

Production

Schema	Production
cdm	21,89%
dds	8,89%
ods	39,25%
raw	25,60%
rep	0,00%
tech	4,29%
udm	0,08%

Team

Select Team to drill down

Team	Share
Atlas	0,01%
Demand	10,20%
Marketplace	30,01%
Supply Eda	34,71%
Supply Taxi	22,65%

User

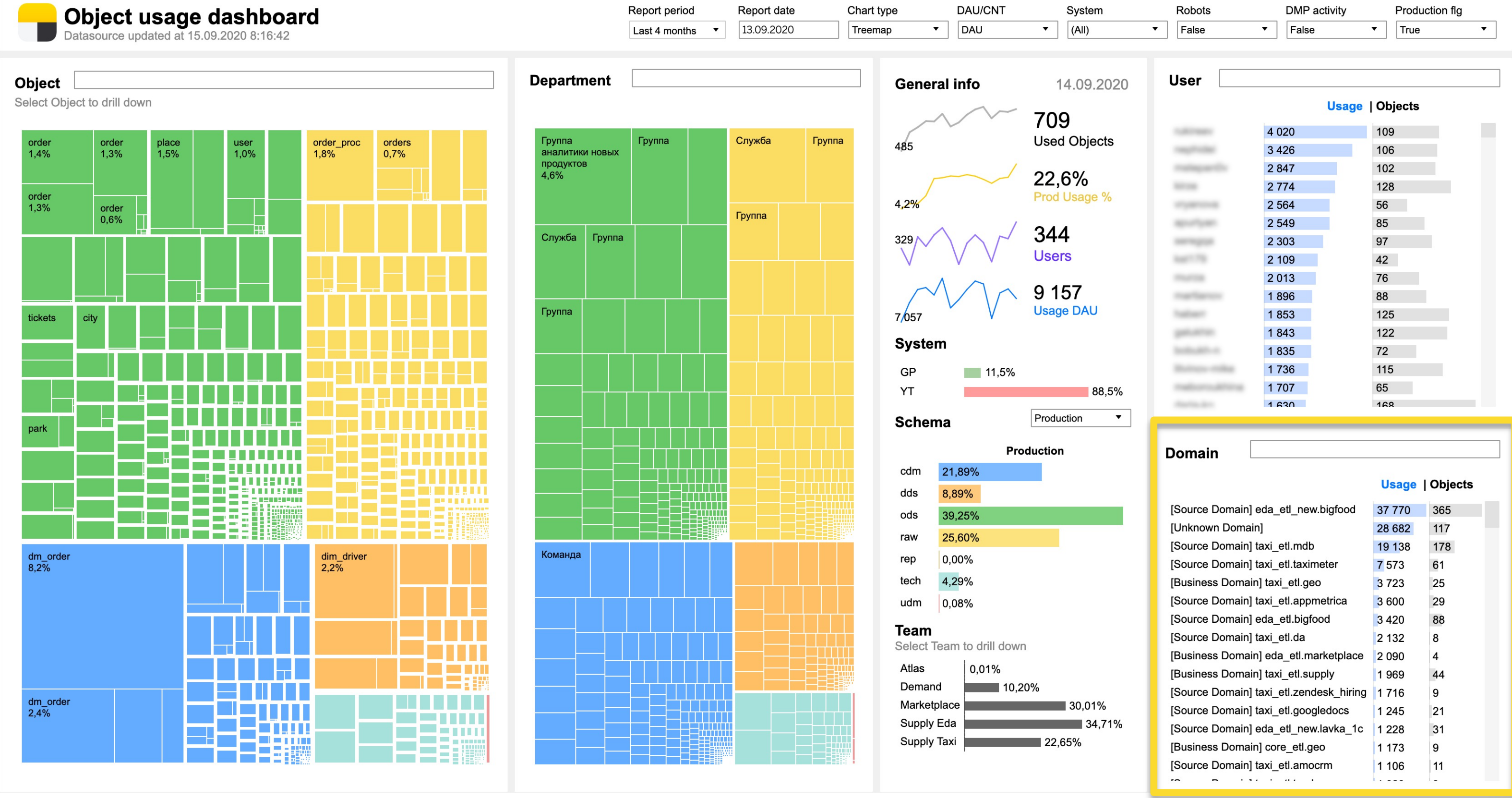
Usage	Objects
4 020	109
3 426	106
2 847	102
2 774	128
2 564	56
2 549	85
2 303	97
2 109	42
2 013	76
1 896	88
1 853	125
1 843	122
1 835	72
1 736	115
1 707	65
1 630	168

Domain

Domain	Usage	Objects
[Source Domain] eda_etl_new.bigfood	37 770	365
[Unknown Domain]	28 682	117
[Source Domain] taxi_etl.mdb	19 138	178
[Source Domain] taxi_etl.taximeter	7 573	61
[Business Domain] taxi_etl.geo	3 723	25
[Source Domain] taxi_etl.appmetrica	3 600	29
[Source Domain] eda_etl.bigfood	3 420	88
[Source Domain] taxi_etl.da	2 132	8
[Business Domain] eda_etl.marketplace	2 090	4
[Business Domain] taxi_etl.supply	1 969	44
[Source Domain] taxi_etl.zendesk_hiring	1 716	9
[Source Domain] taxi_etl.googledocs	1 245	21
[Source Domain] eda_etl_new.lavka_1c	1 228	31
[Business Domain] core_etl.geo	1 173	9
[Source Domain] taxi_etl.amocrm	1 106	11



Использование объектов



Использование объектов

Object usage dashboard

Datasource updated at 15.09.2020 8:16:42

Report period	Report date	Chart type	DAU/CNT	System	Robots	DMP activity	Production flg
Last 4 months	13.09.2020	Treemap	DAU	(All)	False	False	True

Object

Select Object to drill down

order 1,4%
order 1,3%
place 1,5%
user 1,0%
order_proc 1,8%
orders 0,7%
order 1,3%
order 0,6%
tickets
city
park
dm_order 8,2%
dm_driver 2,2%
dm_order 2,4%

Department

Группа аналитики новых продуктов 4,6%
Группа
Служба
Группа
Группа
Служба
Группа
Группа
Команда

General info 14.09.2020

485 — 709 Used Objects

4,2% — 22,6% Prod Usage %

329 — 344 Users

7 057 — 9 157 Usage DAU

System

System	%
GP	11,5%
YT	88,5%

Schema

Production

Schema	Production
cdm	21,89%
dds	8,89%
ods	39,25%
raw	25,60%
rep	0,00%
tech	4,29%
udm	0,08%

Team

Select Team to drill down

Team	%
Atlas	0,01%
Demand	10,20%
Marketplace	30,01%
Supply Eda	34,71%
Supply Taxi	22,65%

User

User	Usage	Objects
	4 020	109
	3 426	106
	2 847	102
	2 774	128
	2 564	56
	2 549	85
	2 303	97
	2 109	42
	2 013	76
	1 896	88
	1 853	125
	1 843	122
	1 835	72
	1 736	115
	1 707	65
	1 630	168

Domain

Domain	Usage	Objects
[Source Domain] eda_etl_new.bigfood	37 770	365
[Unknown Domain]	28 682	117
[Source Domain] taxi_etl.mdb	19 138	178
[Source Domain] taxi_etl.taximeter	7 573	61
[Business Domain] taxi_etl.geo	3 723	25
[Source Domain] taxi_etl.appmetrica	3 600	29
[Source Domain] eda_etl.bigfood	3 420	88
[Source Domain] taxi_etl.da	2 132	8
[Business Domain] eda_etl.marketplace	2 090	4
[Business Domain] taxi_etl.supply	1 969	44
[Source Domain] taxi_etl.zendesk_hiring	1 716	9
[Source Domain] taxi_etl.googledocs	1 245	21
[Source Domain] eda_etl_new.lavka_1c	1 228	31
[Business Domain] core_etl.geo	1 173	9
[Source Domain] taxi_etl.amocrm	1 106	11



Использование объектов

Object usage dashboard
Datasource updated at 15.09.2020 8:16:42

Report period	Report date	Chart type	DAU/CNT	System	Robots	DMP activity	Production flg
Last 4 months	13.09.2020	Treemap	DAU	(All)	False	False	True

Object
Select Object to drill down

order 1,4%
order 1,3%
place 1,5%
user 1,0%
order_proc 1,8%
orders 0,7%
order 1,3%
order 0,6%
tickets
city
park
dm_order 8,2%
dm_driver 2,2%
dm_order 2,4%

Department

Группа аналитики новых продуктов 4,6%
Группа
Служба
Группа
Группа
Служба
Группа
Группа
Команда

General info 14.09.2020

485
709 Used Objects

4,2%
22,6% Prod Usage %

329
344 Users

7 057
9 157 Usage DAU

System

GP	11,5%
YT	88,5%

Schema Production

Production

cdm	21,89%
dds	8,89%
ods	39,25%
raw	25,60%
rep	0,00%
tech	4,29%
udm	0,08%

Team
Select Team to drill down

Atlas	0,01%
Demand	10,20%
Marketplace	30,01%
Supply Eda	34,71%
Supply Taxi	22,65%

User

Usage | Objects

Usage	Objects
4 020	109
3 426	106
2 847	102
2 774	128
2 564	56
2 549	85
2 303	97
2 109	42
2 013	76
1 896	88
1 853	125
1 843	122
1 835	72
1 736	115
1 707	65
1 630	168

Domain

Usage | Objects

Domain	Usage	Objects
[Source Domain] eda_etl_new.bigfood	37 770	365
[Unknown Domain]	28 682	117
[Source Domain] taxi_etl.mdb	19 138	178
[Source Domain] taxi_etl.taximeter	7 573	61
[Business Domain] taxi_etl.geo	3 723	25
[Source Domain] taxi_etl.appmetrica	3 600	29
[Source Domain] eda_etl.bigfood	3 420	88
[Source Domain] taxi_etl.da	2 132	8
[Business Domain] eda_etl.marketplace	2 090	4
[Business Domain] taxi_etl.supply	1 969	44
[Source Domain] taxi_etl.zendesk_hiring	1 716	9
[Source Domain] taxi_etl.googledocs	1 245	21
[Source Domain] eda_etl_new.lavka_1c	1 228	31
[Business Domain] core_etl.geo	1 173	9
[Source Domain] taxi_etl.amocrm	1 106	11



Нотификация про изменения

При изменении объекта мы знаем, кто им пользовался, и можем точечно уведомить про изменения в конкретном объекте в любом из доступных каналов коммуникации.

Пример письма:

Действие:
удалить

Объект:
//home/taxi-dwh/dds/driver_session_geoposition/

Причина действия:
В конце февраля 2021 мы остановим загрузку и перестанем поддерживать объект dds.driver_session_geoposition

Альтернативные объекты:
fct_supply_state_hist - водительские сессии, в которых собрано большинство атрибутов по активности водителей
fct_taxi_tracker_position_enriched_log - замена driver_session_geoposition, построенная на логе водительских геопозиций и fct_supply_state_hist

Пожалуйста, переведите ваши процессы на новые объекты и сообщите нам о сроках, когда вы сможете запланировать переезд.

Вики как переезжать - https://wiki.yandex-team.ru/taxi/dwh/data/business/driver-session/kak-perejjti-s-driversession-na-fctsupplystatehist/

Удаление запланировано на конец февраля 2021

Ссылка на тикет, в котором мы ведём работу над удалением/изменением:
TAXIDWH-5913 Открыт Удалить driver_session_geoposition avbekker

В этот день мы удалим/изменим объект:
2021-02-28



III. Что получили?

КАК ОЦЕНИТЬ РАБОТУ ПРОДУКТОВОЙ КОМАНДЫ?



Что нам важно?

Результатом нашей работы (Объектами DWH) пользуются

Считаем уникальные пары (пользователь, используемый DWH-объект) за каждый день и убираем те объекты, которыми пользуется ровно один пользователь (его личная песочница). На результирующих данных можем посчитать:



Что нам важно?

Результатом нашей работы (Объектами DWH) пользуются

Считаем уникальные пары (пользователь, используемый DWH-объект) за каждый день и убираем те объекты, которыми пользуется ровно один пользователь (его личная песочница). На результирующих данных можем посчитать:

- соотношение обращений к prod- и не prod-объектам
  показывает, насколько пользователи смотрят в prod объекты

- соотношение обращений к deprecated- и не deprecated-объектам
  показывает, насколько мы избавляемся от легаси

- соотношение обращений tech vs all
  показывает, насколько мы быстро расшифровываем новые данные

- соотношение обращений CDM+REP vs all
  показывает, насколько наши целевые объекты удобны пользователям



Количественные KPI команд

DMP - Usage KPI - All teams

Datasource updated at 09.05.2021 8:57:02

Report period	Report date	DAU/CNT	System	Robots	DMP activity	Team
Last 13 months	18.04.2021	CNT	(All)	False	False	All teams

Usage by layer - All teams

Select Layer to drill down

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

	Production	Not Production
cdm	19,38%	0,37%
dds	0,98%	0,27%
ods	21,86%	0,24%
raw	8,61%	0,40%
rep	0,02%	0,00%
snb	0,00%	44,25%
tech	3,44%	0,17%
udm	0,02%	0,00%

Production Kpi

Select KPI to filter

25,6%

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

54,31%
▲4,4% MoM

GP 15,75%

YT 38,55%

Team Kpi - All teams

Select KPI to filter

CDM / REP or Summary

7,3%

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

19,77%
▲0,1% MoM

Not Deprecated

76,7%

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

79,34%
▲0,7% MoM

Not Tech (RAW / STG)

90,0%

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

87,38%
▼1,6% MoM

Production by team

Select Team to filter

All teams

Atlas

Demand

Marketplace

MassService

Supply Eda

Supply Taxi

Undefined

Production - Object

Select Object to drill down

Object / Dep / User: Object

DAU/CNT at least: 200

YT	dm_order cdm [dep]	758 697
YT	quality_control_log raw	206 596
YT	driver_session_reduced dds [dep]	187 349
YT	dim_driver dds [dep]	146 763
YT	pin_stats raw	142 937
YT	dm_order cdm	136 843
YT	user_phone ods	112 053
YT	order_proc raw	106 061

Not Production - Object

YT	disinfected_cars_table_ snb	120 584
GP	voytekh_manager_v2 snb [dep]	115 624
GP	voytekh_pd_user snb [dep]	102 287
YT	accra_branding_subs snb	100 995
YT	helsinki_week snb	100 847
YT	support_tickets cdm [dep]	96 051



Количественные KPI команд

DMP - Usage KPI - All teams

Datasource updated at 09.05.2021 8:57:02

Report period	Report date	DAU/CNT	System	Robots	DMP activity	Team
Last 13 months	18.04.2021	CNT	(All)	False	False	All teams

Usage by layer - All teams

Select Layer to drill down

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

	Production	Not Production
cdm	19,38%	0,37%
dds	0,98%	0,27%
ods	21,86%	0,24%
raw	8,61%	0,40%
rep	0,02%	0,00%
snb	0,00%	44,25%
tech	3,44%	0,17%
udm	0,02%	0,00%

Production Kpi

Select KPI to filter

25,6%

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

54,31%
▲4,4% MoM
GP 15,75%
YT 38,55%

Team Kpi - All teams

Select KPI to filter

CDM / REP or Summary

7,3%

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

19,77%
▲0,1% MoM

Not Deprecated

76,7%

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

79,34%
▲0,7% MoM

Not Tech (RAW / STG)

90,0%

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

87,38%
▼1,6% MoM

Production by team

Select Team to filter

All teams
Atlas
Demand
Marketplace
MassService
Supply Eda
Supply Taxi
Undefined

Production - Object

Select Object to drill down

Object / Dep / User: Object
DAU/CNT at least: 200

YT	dm_order cdm [dep]	758 697
YT	quality_control_log raw	206 596
YT	driver_session_reduced dds [dep]	187 349
YT	dim_driver dds [dep]	146 763
YT	pin_stats raw	142 937
YT	dm_order cdm	136 843
YT	user_phone ods	112 053
YT	order_proc raw	106 061

Not Production - Object

YT	disinfected_cars_table_ snb	120 584
GP	voytekh_manager_v2 snb [dep]	115 624
GP	voytekh_pd_user snb [dep]	102 287
YT	accra_branding_subs snb	100 995
YT	helsinki_week snb	100 847
YT	support_tickets cdm [dep]	96 051



Количественные KPI команд

DMP - Usage KPI - All teams

Datasource updated at 09.05.2021 8:57:02

Report period	Report date	DAU/CNT	System	Robots	DMP activity	Team
Last 13 months	18.04.2021	CNT	(All)	False	False	All teams

Usage by layer - All teams

Select Layer to drill down

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

	Production	Not Production
cdm	19,38%	0,37%
dds	0,98%	0,27%
ods	21,86%	0,24%
raw	8,61%	0,40%
rep	0,02%	0,00%
snb	0,00%	44,25%
tech	3,44%	0,17%
udm	0,02%	0,00%

Production Kpi

Select KPI to filter

25,6%

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

54,31%
▲4,4% MoM

GP 15,75%
YT 38,55%

Team Kpi - All teams

Select KPI to filter

CDM / REP or Summary

7,3%

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

19,77%
▲0,1% MoM

Not Deprecated

76,7%

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

79,34%
▲0,7% MoM

Not Tech (RAW / STG)

90,0%

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

87,38%
▼1,6% MoM

Production by team

Select Team to filter

All teams
Atlas
Demand
Marketplace
MassService
Supply Eda
Supply Taxi
Undefined

Production - Object

Select Object to drill down

Object / Dep / User: Object
DAU/CNT at least: 200

YT	dm_order cdm [dep]	758 697
YT	quality_control_log raw	206 596
YT	driver_session_reduced dds [dep]	187 349
YT	dim_driver dds [dep]	146 763
YT	pin_stats raw	142 937
YT	dm_order cdm	136 843
YT	user_phone ods	112 053
YT	order_proc raw	106 061

Not Production - Object

YT	disinfected_cars_table_ snb	120 584
GP	voytekh_manager_v2 snb [dep]	115 624
GP	voytekh_pd_user snb [dep]	102 287
YT	accra_branding_subs snb	100 995
YT	helsinki_week snb	100 847
YT	support_tickets cdm [dep]	96 051



Количественные KPI команд

DMP - Usage KPI - All teams

Datasource updated at 09.05.2021 8:57:02

Report period	Report date	DAU/CNT	System	Robots	DMP activity	Team
Last 13 months	18.04.2021	CNT	(All)	False	False	All teams

Usage by layer - All teams

Select Layer to drill down

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

	Production	Not Production
cdm	19,38%	0,37%
dds	0,98%	0,27%
ods	21,86%	0,24%
raw	8,61%	0,40%
rep	0,02%	0,00%
snb	0,00%	44,25%
tech	3,44%	0,17%
udm	0,02%	0,00%

Production Kpi

Select KPI to filter

25,6%

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

54,31%

▲4,4% MoM

GP 15,75%

YT 38,55%

Team Kpi - All teams

Select KPI to filter

CDM / REP or Summary

7,3%

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

19,77%

▲0,1% MoM

Not Deprecated

76,7%

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

79,34%

▲0,7% MoM

Not Tech (RAW / STG)

90,0%

6 июля 20 г. 28 сент. 20 г. 21 дек. 20 г. 15 марта 21 г.

87,38%

▼1,6% MoM

Production by team

Select Team to filter

- All teams
- Atlas
- Demand
- Marketplace
- MassService
- Supply Eda
- Supply Taxi
- Undefined

Production - Object

Select Object to drill down

Object / Dep / User: Object

DAU/CNT at least: 200

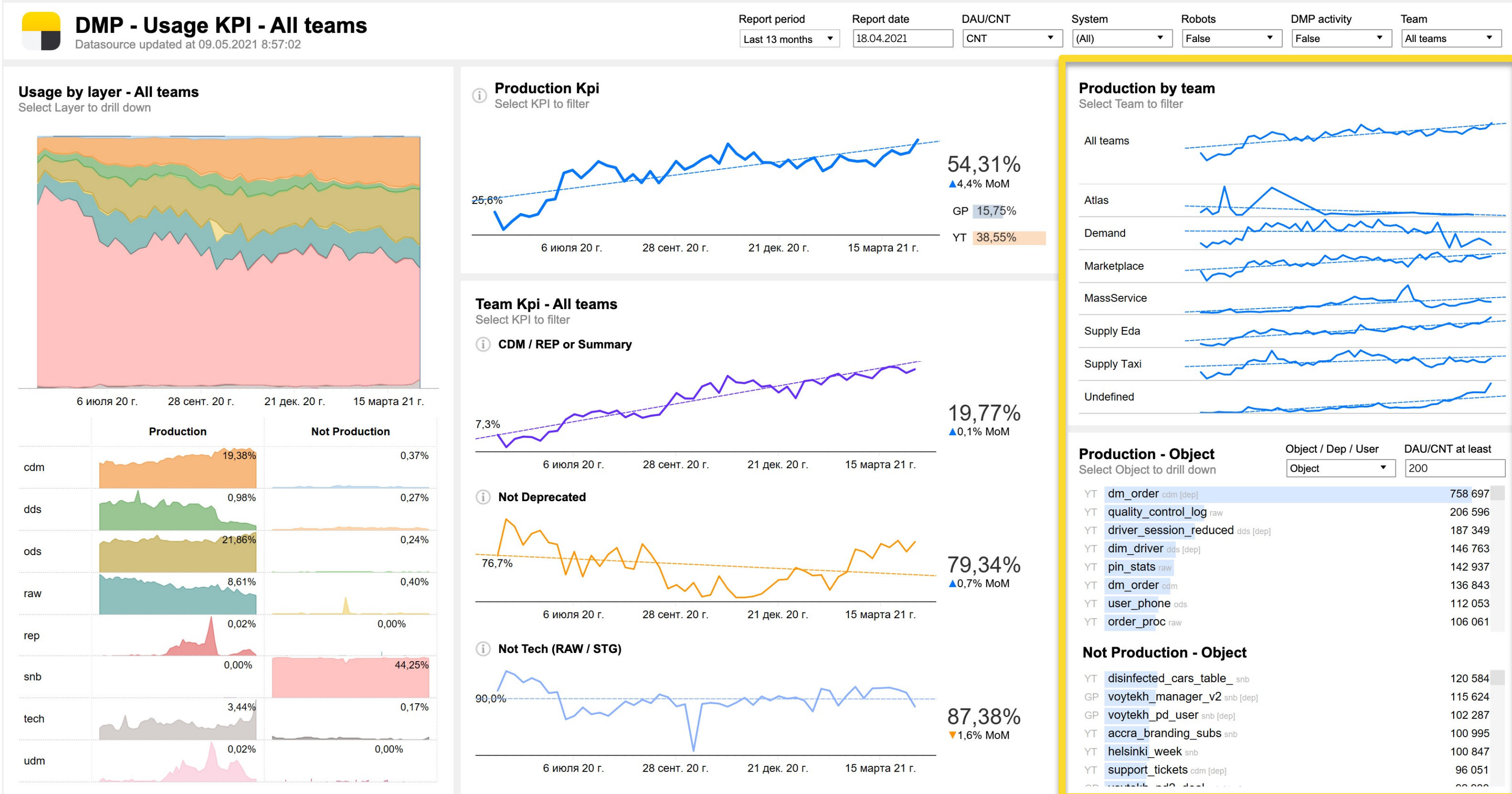
YT	dm_order cdm [dep]	758 697
YT	quality_control_log raw	206 596
YT	driver_session_reduced dds [dep]	187 349
YT	dim_driver dds [dep]	146 763
YT	pin_stats raw	142 937
YT	dm_order cdm	136 843
YT	user_phone ods	112 053
YT	order_proc raw	106 061

Not Production - Object

YT	disinfected_cars_table_ snb	120 584
GP	voytekh_manager_v2 snb [dep]	115 624
GP	voytekh_pd_user snb [dep]	102 287
YT	accra_branding_subs snb	100 995
YT	helsinki_week snb	100 847
YT	support_tickets cdm [dep]	96 051



Количественные KPI команд



Оценка качества домена

Можем ввести метрики, косвенно оценивающие качество доменов

Архитектура

- Соблюдение naming convention
- Использование legacy-объектов
- Доля витрин, построенных на базе source domain (RAW|ODS), а не core domain (DDS|CDM)

Качество данных

- Отсутствие ПД
- Скорость поставки данных
- Качество документации
- Покрытие данных проверками качества

Качество расчетов

- Использование последних инструментов платформы
- Оптимальность ETL-процессов
- Недоступность (downtime) объектов

Итоговая оценка качества домена как взвешенная сумма критериев



III. Что получили?

ЕСТЬ ЛИ БОЛЕЕ ТЕХНИЧЕСКОЕ ПРИМЕНЕНИЕ?



Размер объектов DWH

YT Resource usage
Last datasource update at 09.05.2021 12:47:34

Creation time: 06.10.20 – 09.05.202
Modify time: 06.10.20 – 09.05.202
Measure type: Data Weight
Level 0: taxi-dwh
Erasure Codec: (All)
Compression Codec: (All)

Resource usage by Data Weight
Color is level 1

appmetrica 684 113 972
experiments 115 734 150
mdb 57 126 511
raw_history 362 342 619
mdb 248 910 360
dbtaxi 76 924 184
food 55 732 900
ab_experiment 75 539 392
demand 54 427 158
dbprocessing 184 822 097
taximeter 38 861 969
adjust
tracker 33 312 233
xiva

Data Weight
2 369 506 829M

Level 1

ods	975 823 484M
raw	719 269 322M
etl	364 409 059M
cdm	151 896 787M
dds	105 591 753M
import	24 937 593M
summary	12 037 387M
export	5 136 157M
backup	4 061 661M
stg	3 943 135M
public	2 281 980M

Level 3

taximeter	461 396 626M
orders	354 550 991M
order_proc	184 822 097M
taxi	174 292 275M
dbtaxi	157 843 214M
experiment_v3_log	115 733 676M
fct_experiment_v3_log	75 539 389M
road_accidents	72 328 976M
driver_position	56 328 114M
adjust	55 129 678M
place_surge_log	48 618 928M
orders_monthly	33 481 446M
pin_stats	33 006 566M
eventlog	29 190 059M
order	25 890 423M
orders_raw	21 075 192M
user_appsession	20 533 601M

Level 2

appmetrica	684 244 878M
raw_history	362 342 619M
mdb	309 799 824M
dbprocessing	205 020 577M
experiments	116 511 055M
dbtaxi	80 913 305M
ab_experiment	75 539 392M
driver_session_...	67 389 687M
food	58 429 652M
tracker	56 331 193M
demand	54 434 874M
eda_prod_surge	48 618 928M
taximeter	40 323 583M
adjust	29 181 108M
driver_session	24 527 922M
supply	18 543 549M
TAXIDWH-101..	12 383 775M

Level 4

order_proc	180 633 251M
orders	153 897 051M
	142 725 172M
2021-04	22 698 217M
2021-04-01	22 250 300M
2020-10-01	21 744 922M
2021-03-01	21 554 095M
2020-11-01	20 724 729M
2020-09-01	20 504 719M
2021-03	19 585 367M
2020-12-01	18 752 159M
2020-08-01	18 530 263M
2021-02-01	17 511 327M
2021-01-01	17 359 758M
2020-07-01	16 876 562M
2020-12	16 447 576M
2021-02	16 218 177M



Потребление ресурсов

Task resource usage dashboard
Datasource updated at 20.04.2021 7:31:13

Month: Апрель 2021
KPI: AVG Run, sec
Task Name: (All)

AVG Run, sec by task
Color is Layer

general_user_metrics
80 997

yt_load_lavka_user_appsession
64 762

public_maps_order
24 078

stg_device_taxi_profile
46 577

AVG Run, sec

1 104,92
1 457,49
13 сент. 20 г.
14 марта 21 г.

Layer

REP	1 861
CDM	2 652
DDS	2 477
STG	803
ODS	1 010
RAW	444
More than 1 layer	3 960

Business unit

Taxi	1 873
Eda	784
Lavka	1 821
More than 1 unit	775
Undefined	1 625

System

GP	842
YT	2 746
More than 1 system	470
Undefined	1 515



III. Что получили?

МОЖНО ЛИ ПРИМЕНИТЬ ЗНАНИЯ ДЛЯ ОПТИМИЗАЦИИ?



Детальный слой

Детальный слой – ключевой для построения доменной модели

- Хранить историю изменений сущностей
- Отвечает за консолидацию данных между источниками
- Устойчив к изменению в бизнесе
- Модульный и масштабируемый

Greenplum

DDS

Николай Гребенщиков, Ермаков Евгений,
https://smartdataconf.ru/2020/spb/talks/1o78kcrupy6qcaaff4m5oh/



Подходы к проектированию

сложность эксплуатации, простота внесения изменений →

Никакого

- Денормализация
- Можно использовать без подготовки
- Неустойчиво к изменениям
- Дублирование информации
- Нет join

Звезда и снежинка

- Нормализация
- Можно использовать с минимальной подготовкой
- Неудобно перестраивать
- Минимальное дублирование информации
- Приемлемое количество join

Data Vault

- Строгая нормализация
- Нельзя использовать без подготовки
- Не надо перестраивать
- Нет дублирования информации
- Большое количество join

Anchor modeling

- Ультранормализация
- Нельзя использовать без подготовки
- Не надо перестраивать
- Нет дублирования информации
- Ультраколичество join

← легкость эксплуатации, сложность внесения изменений

Николай Гребенщиков, Ермаков Евгений,
https://smartdataconf.ru/2020/spb/talks/1o78kcrupy6qcaaff4m5oh/



Highly Normalized Hybrid Model

Выбирать оптимальный формат хранения для каждого конкретного случая

- Высокая нормализация
- Параллельная загрузка из разных источников
- Устойчив к изменению в бизнесе
- Идемпотентный к повторной загрузке
- Модульный и масштабируемый
- Может эмулировать как Data Vault, так и Anchor Modeling



- Атрибуты группируются в таблицы-**сателлиты** по принципам совместности: изменения и/или источника и/или использования
- Есть специальные таблицы **Point-in-Time** и **Bridge**



- На каждую сущность создается **anchor** – таблица с суррогатным ключом
- **Связи** только через отдельные таблицы, никаких атрибутов – только хардкор

Николай Гребенщиков, Ермаков Евгений, https://smartdataconf.ru/2020/spb/talks/1o78kcrupy6qcaaff4m5oh/



Объявление сущности

```
class Person(HnhmEntity):
    """Сотрудник со staff.yandex-team.ru"""

    __layout__ = DdsLayout(name='person', group='staff')

    person_id        = Int(comment='ID в Стаффе', change_type=IGNORE)
    first_name_ru    = String(comment='Имя сотрудника', change_type=UPDATE)
    last_name_ru     = String(comment='Фамилия сотрудника', change_type=NEW)
    login            = String(comment='Рабочий login', change_type=IGNORE)
    gender           = String(comment='Пол', change_type=UPDATE)
    tshirt_size      = String(comment='Размер футболки', change_type=UPDATE)
    birthday_dt      = Date(comment='Дата рождения', change_type=UPDATE)
    is_dismissed_flg  = Boolean(comment='Был уволен',change_type=NEW)
   is_homeworker_flg  = Boolean(comment='Надомник',change_type=NEW)
   is_robot_flg      = Boolean(comment='Робот',change_type=NEW)
    is_trainee_flg   = Boolean(comment='Стажер',change_type=NEW)

    __keys__ = [login]
```




a__person__person_id	
PK	person_sk
	person_id
	utc_valid_from_dttm
	_utc_etl_processed_dttm
	_source_id

a__person__first_name_ru	
PK	person_sk
	first_name_ru
	utc_valid_from_dttm
	_utc_etl_processed_dttm
	_source_id

a__person__is_homeworker_flg	
PK	person_sk
	is_homeworker_flg
	utc_valid_from_dttm
	utc_valid_to_dttm
	_utc_etl_processed_dttm
	_source_id

a__person__is_dismissed_flg	
PK	person_sk
	is_dissmissed_flg
	utc_valid_from_dttm
	utc_valid_to_dttm
	_utc_etl_processed_dttm
	_source_id

a__person__is_trainee_flg	
PK	person_sk
	is_trainee_flg
	utc_valid_from_dttm
	utc_valid_to_dttm
	_utc_etl_processed_dttm
	_source_id

h__person	
PK	person_sk
	utc_valid_from_dttm
	_utc_etl_processed_dttm
	_source_id

a__person__last_name_ru	
PK	person_sk
	last_name_ru
	utc_valid_from_dttm
	utc_valid_to_dttm
	_utc_etl_processed_dttm
	_source_id

a__person__is_robot_flg	
PK	person_sk
	is_robot_flg
	utc_valid_from_dttm
	utc_valid_to_dttm
	_utc_etl_processed_dttm
	_source_id

a__person__tshirt_size	
PK	person_sk
	tshirt_size
	utc_valid_from_dttm
	_utc_etl_processed_dttm
	_source_id

a__person__birthday_dt	
PK	person_sk
	birthday_dt
	utc_valid_from_dttm
	_utc_etl_processed_dttm
	_source_id

a__person__gender	
PK	person_sk
	gender
	utc_valid_from_dttm
	_utc_etl_processed_dttm
	_source_id

a__person__login__key	
PK	person_sk
	login
	utc_valid_from_dttm
	_utc_etl_processed_dttm
	_source_id



g__person__flg	
PK	person_sk
	is_homeworker_flg
	is_robot_flg
	is_dissmissed_flg
	is_trainee_flg
	utc_valid_from_dttm
	utc_valid_to_dttm
	_utc_etl_processed_dttm
	_source_id

a__person__first_name_ru	
PK	person_sk
	first_name_ru
	utc_valid_from_dttm
	_utc_etl_processed_dttm
	_source_id

a__person__last_name_ru	
PK	person_sk
	last_name_ru
	utc_valid_from_dttm
	utc_valid_to_dttm
	_utc_etl_processed_dttm
	_source_id

h__person	
PK	person_sk
	utc_valid_from_dttm
	_utc_etl_processed_dttm
	_source_id

g__person__key	
PK	person_sk
	login
	person_id
	utc_valid_from_dttm
	_utc_etl_processed_dttm
	_source_id

g__person__info	
PK	person_sk
	birthday_dt
	gender
	tshirt_size
	utc_valid_from_dttm
	utc_valid_to_dttm
	_utc_etl_processed_dttm
	_source_id



Оптимизационная задача

Вопрос: как оптимально разбить данные по группам?

Дано (и есть в metaDWH):

- Метаданные объектов
- Маппинги полей и загрузчики
- Информация о количестве строк в объекте

Ограничения

- Набор полей в метаданных объектов
- Маппинги полей и загрузчики (группа должна загружаться из одного источника

Оптимальность

- Будем минимизировать занимаемое место на диске



Синтетический пример

business_dttm
key
Field 1
Field 2
Field 3

Сущность состоит из 3х полей

- Key – 16 байт
- Field1 – 8байт
- Field2 – 128байт
- Fileld3 – 32байт

Итого 192 байт на строку

Дополнительные поля

- _sk – 16 байт
- _dttm – 8 байт
- _source_id – 2байт

Итого 42 байт на строку



Синтетический пример

business_dttm
key
Field 1
Field 2
Field 3

Сущность состоит из 3х полей

- Key – 16 байт
- Field1 – 8байт
- Field2 – 128байт
- Fileld3 – 32байт

Итого 192 байт на строку

Дополнительные поля

- _sk – 16 байт
- _dttm – 8 байт
- _source_id – 2байт

Итого 42 байт на строку

F1F2F3

sk
Field 1
Field 2
Field 3
valid_from_dttm
valid_to_dttm
_source_id
_etl_dttm

F1 | F2F3

sk	sk
Field 1	Field 2
valid_from_dttm	Field 3
valid_to_dttm	valid_from_dttm
_source_id	valid_to_dttm
_etl_dttm	_source_id
	_etl_dttm

F2 | F1F3

sk	sk
Field 2	Field 1
valid_from_dttm	Field 3
valid_to_dttm	valid_from_dttm
_source_id	valid_to_dttm
_etl_dttm	_source_id
	_etl_dttm

F3 | F1F2

sk	sk
Field 3	Field 2
valid_from_dttm	Field 1
valid_to_dttm	valid_from_dttm
_source_id	valid_to_dttm
_etl_dttm	_source_id
	_etl_dttm

F1 | F1 | F3

sk	sk	sk
Field 1	Field 1	Field 1
valid_from_dttm	valid_from_dttm	valid_from_dttm
valid_to_dttm	valid_to_dttm	valid_to_dttm
_source_id	_source_id	_source_id
_etl_dttm	_etl_dttm	_etl_dttm



Синтетический пример

business_dttm
key
Field 1
Field 2
Field 3

Сущность состоит из 3х полей – 1 000 000 строчек

- Key – 16 байт – 500 000 сущностей
- Field1 – 8байт – 2 изменения на ключ
- Field2 – 128байт – 2 изменения на ключ
- Fileld3 – 32байт – 2 изменения на ключ

Итого 183 Мбайт

Дополнительные поля

- _sk – 16 байт
- _dttm – 8 байт
- _source_id – 2байт

Итого 42 байт

F1F2F3

sk
Field 1
Field 2
Field 3
valid_from_dttm
valid_to_dttm
_source_id
_etl_dttm

200 Мбайт

F1 | F2F3

sk	sk
Field 1	Field 2
valid_from_dttm	Field 3
valid_to_dttm	valid_from_dttm
_source_id	valid_to_dttm
_etl_dttm	_source_id
	_etl_dttm

240 Мбайт

F2 | F1F3

sk	sk
Field 2	Field 1
valid_from_dttm	Field 3
valid_to_dttm	valid_from_dttm
_source_id	valid_to_dttm
_etl_dttm	_source_id
	_etl_dttm

240 Мбайт

F3 | F1F2

sk	sk
Field 3	Field 2
valid_from_dttm	Field 1
valid_to_dttm	valid_from_dttm
_source_id	valid_to_dttm
_etl_dttm	_source_id
	_etl_dttm

240 Мбайт

F1 | F1 | F3

sk	sk	sk
Field 1	Field 1	Field 1
valid_from_dttm	valid_from_dttm	valid_from_dttm
valid_to_dttm	valid_to_dttm	valid_to_dttm
_source_id	_source_id	_source_id
_etl_dttm	_etl_dttm	_etl_dttm

280 Мбайт

Синтетический пример

business_dttm
key
Field 1
Field 2
Field 3

Сущность состоит из 3х полей – 1 000 000 строчек

- Key – 16 байт – 31 250 сущностей
- Field1 – 8байт – 2 изменения на ключ
- Field2 – 128байт – **32 изменения** на ключ
- Fileld3 – 32байт – 2 изменения на ключ

Итого 183 Мбайт

Дополнительные поля

- _sk – 16 байт
- _dttm – 8 байт
- _source_id – 2байт

Итого 42 байт

F1F2F3

sk
Field 1
Field 2
Field 3
valid_from_dttm
valid_to_dttm
_source_id
_etl_dttm

200 Мбайт

F1 | F2F3

sk
Field 1
valid_from_dttm
valid_to_dttm
_source_id
_etl_dttm

sk
Field 2
Field 3
valid_from_dttm
valid_to_dttm
_source_id
_etl_dttm

174 Мбайт

F2 | F1F3

sk
Field 2
valid_from_dttm
valid_to_dttm
_source_id
_etl_dttm

sk
Field 1
Field 3
valid_from_dttm
valid_to_dttm
_source_id
_etl_dttm

167 Мбайт

F3 | F1F2

sk
Field 3
valid_from_dttm
valid_to_dttm
_source_id
_etl_dttm

sk
Field 2
Field 1
valid_from_dttm
valid_to_dttm
_source_id
_etl_dttm

195 Мбайт

F1 | F1 | F3

sk
Field 1
valid_from_dttm
valid_to_dttm
_source_id
_etl_dttm

sk
Field 1
valid_from_dttm
valid_to_dttm
_source_id
_etl_dttm

sk
Field 1
valid_from_dttm
valid_to_dttm
_source_id
_etl_dttm

169 Мбайт

Синтетический пример

business_dttm
key
Field 1
Field 2
Field 3

Сущность состоит из 3х полей – 1 000 000 строчек

- Key – 16 байт – 15 625 сущностей
- Field1 – 8байт – 2 изменения на ключ
- Field2 – 128байт – **32 изменения** на ключ
- Fileld3 – 32байт – **64 изменения** на ключ

Итого 183 Мбайт

Дополнительные поля

- _sk – 16 байт
- _dttm – 8 байт
- _source_id – 2байт

Итого 42 байт

F1F2F3

sk
Field 1
Field 2
Field 3
valid_from_dttm
valid_to_dttm
_source_id
_etl_dttm

200 Мбайт

F1 | F2F3

sk
Field 1
valid_from_dttm
valid_to_dttm
_source_id
_etl_dttm

sk
Field 2
Field 3
valid_from_dttm
valid_to_dttm
_source_id
_etl_dttm

155 Мбайт

F2 | F1F3

sk
Field 2
valid_from_dttm
valid_to_dttm
_source_id
_etl_dttm

sk
Field 1
Field 3
valid_from_dttm
valid_to_dttm
_source_id
_etl_dttm

159 Мбайт

F3 | F1F2

sk
Field 3
valid_from_dttm
valid_to_dttm
_source_id
_etl_dttm

sk
Field 2
Field 1
valid_from_dttm
valid_to_dttm
_source_id
_etl_dttm

194 Мбайт

F1 | F1 | F3

sk
Field 1
valid_from_dttm
valid_to_dttm
_source_id
_etl_dttm

sk
Field 1
valid_from_dttm
valid_to_dttm
_source_id
_etl_dttm

sk
Field 1
valid_from_dttm
valid_to_dttm
_source_id
_etl_dttm

150 Мбайт



Наше решение

Вводим атомарные операции, меняющие схему,

но не меняющие логику

- Объединение групп/атрибутов
- Соединение групп/атрибутов

В hNhM с точки зрения использования сущности логической модели все варианты физического хранения ниже одинаковы.

F1F2F3

sk
Field 1
Field 2
Field 3
valid_from_dttm
valid_to_dttm
_source_id
_etl_dttm

F1 | F2F3

sk	sk
Field 1	Field 2
valid_from_dttm	Field 3
valid_to_dttm	valid_from_dttm
_source_id	valid_to_dttm
_etl_dttm	_source_id
	_etl_dttm

F2 | F1F3

sk	sk
Field 2	Field 1
valid_from_dttm	Field 3
valid_to_dttm	valid_from_dttm
_source_id	valid_to_dttm
_etl_dttm	_source_id
	_etl_dttm

F3 | F1F2

sk	sk
Field 3	Field 2
valid_from_dttm	Field 1
valid_to_dttm	valid_from_dttm
_source_id	valid_to_dttm
_etl_dttm	_source_id
	_etl_dttm

F1 | F1 | F3

sk	sk	sk
Field 1	Field 1	Field 1
valid_from_dttm	valid_from_dttm	valid_from_dttm
valid_to_dttm	valid_to_dttm	valid_to_dttm
_source_id	_source_id	_source_id
_etl_dttm	_etl_dttm	_etl_dttm

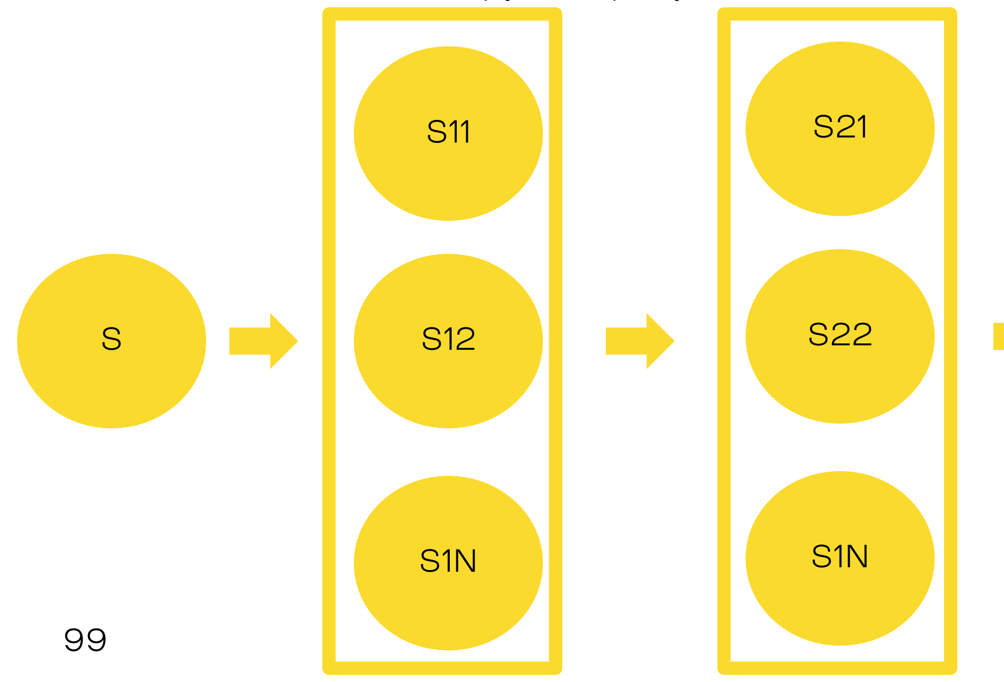


Наше решение

Вводим атомарные операции, меняющие схему,

но не меняющие логику

- Объединение групп/атрибутов
- Соединение групп/атрибутов



Генетический алгоритм

- Из текущего состояния мутациями (=атомарными операциями) создаем стартовую популяцию
- Производим скрещивания и новые мутации
- Каждое состояние оцениваем на оптимальность (в нашем случае по месту)
- При подозрениях на сходимость останавливаемся

Результат

- Получаем итоговое состояние, которое лучше текущего
- Сравниваем метаданные между состояниями и генерируем скрипт миграции
- Миграция – отдельный вопрос



I. ЗАЧЕМ DWH МЕТРИКИ?

II. КАК РЕАЛИЗОВАЛИ?

III. ЧТО ПОЛУЧИЛИ?

IV. СТОИЛО ЛИ ТОГО?



I. Проблема:

развитием крупного DWH сложно управлять

II. Решение:

покрыть работу DWH метриками

III. Идея:

использовать данные систем DWH в самом DWH («DWH для DWH»)

IV. Результат:

аналитика по работе и развитию самого DWH



MetaDWH

Source Domain

Greenplum
- Логи использования
- Метаданные объектов

YT
- Логи использования
- Метаданные объектов

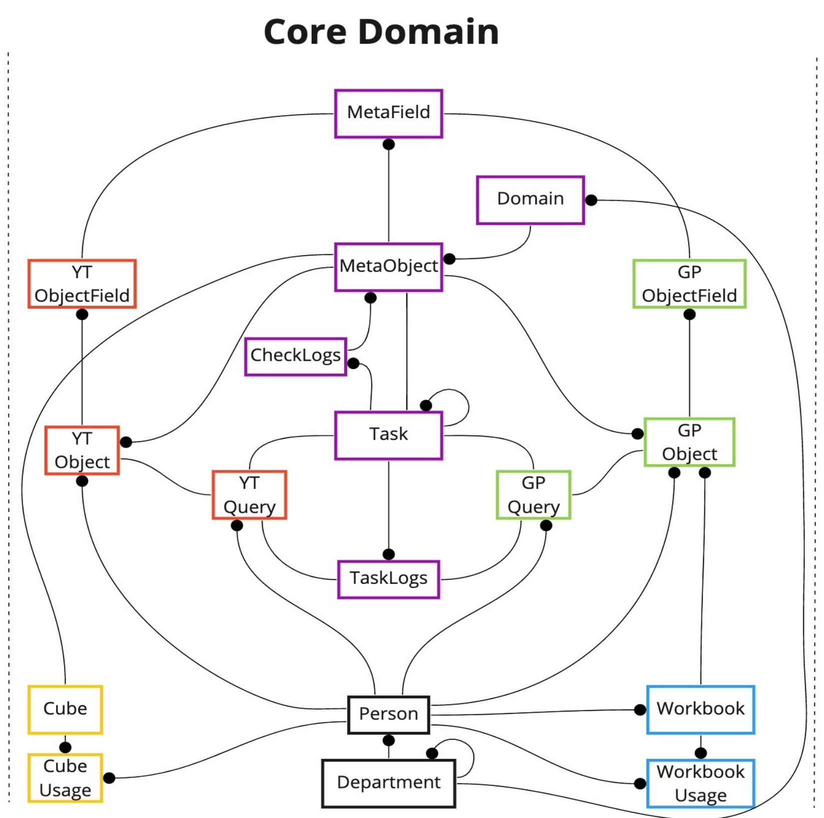
Tableau
- Логи использования
- Метаданные объектов

MS SSAS
- Логи использования
- Метаданные объектов

Staff
- Профиль пользователя
- Орг.Структура

Platform
- Метаданные тасок
- Логи запусков
- Логи проверок
- Метаданные объектов

Core Domain



Business Domain

Техническая информация
- Витрина по размеру данных
- Витрина по потреблению ресурсов

Использование объектов
- Витрина по использованию объектов
- Витрина по использованию отчетов

Витрины с метаданными
- Метаданные объектов
- Витрина по таскам
- Витрина по результатам проверок



Затраты

Стоимость реализации

- Никаких дополнительных внедрений, исключительно существующие системы
- Необходимы разноплановые специалисты (infra, de, dp, bi), которые уже есть в DWH
- Абстрактно в вакууме 2FTE на Q
- Фактически порядка 10 специалистов с ±20% загрузкой



Результат

Стоимость реализации

- Никаких дополнительных внедрений, исключительно существующие системы
- Необходимы разноплановые специалисты (infra, de, dp, bi), которые уже есть в DWH
- Абстрактно в вакууме 2FTE на Q
- Фактически порядка 10 специалистов с ±20% загрузкой

Аналитика по ключевым аспектам

- Целевые метрики и принятие стратегических решений
- управление приоритизацией через KPI команд
- Ad-hoc-запросы по использованию объектов хранилища
- Поиск технически узких мест и оптимизация
- Интеллектуальная нотификация пользователей



Результат

Стоимость реализации

- Никаких дополнительных внедрений, исключительно существующие системы
- Необходимы разноплановые специалисты (infra, de, dp, bi), которые уже есть в DWH
- Абстрактно в вакууме 2FTE на Q
- Фактически порядка 10 специалистов с ±20% загрузкой

<

Аналитика по ключевым аспектам

- Целевые метрики и принятие стратегических решений
- управление приоритизацией через KPI команд
- Ad-hoc-запросы по использованию объектов хранилища
- Поиск технически узких мест и оптимизация
- Интеллектуальная нотификация пользователей

Возможно реализовать на любом отлаженном DWH



Резюме

- DWH может быть источником данных для DWH
- Создать MetaDWH – не слишком трудоемкая задача (при наличии рабочего DWH)
- Обработка только логов запросов позволяет получить дашборды для анализа поведения пользователей
- Более сложная систематизация (домены, слои, команды) позволяет ставить продуктовые метрики командам
- Пример технической реализации: поиск узких мест среди объектов/тасок и модификация схемы в детальном слое



Ермаков Евгений, Яндекс Go

СПАСИБО

https://t.me/iJKos

iJKos.com

